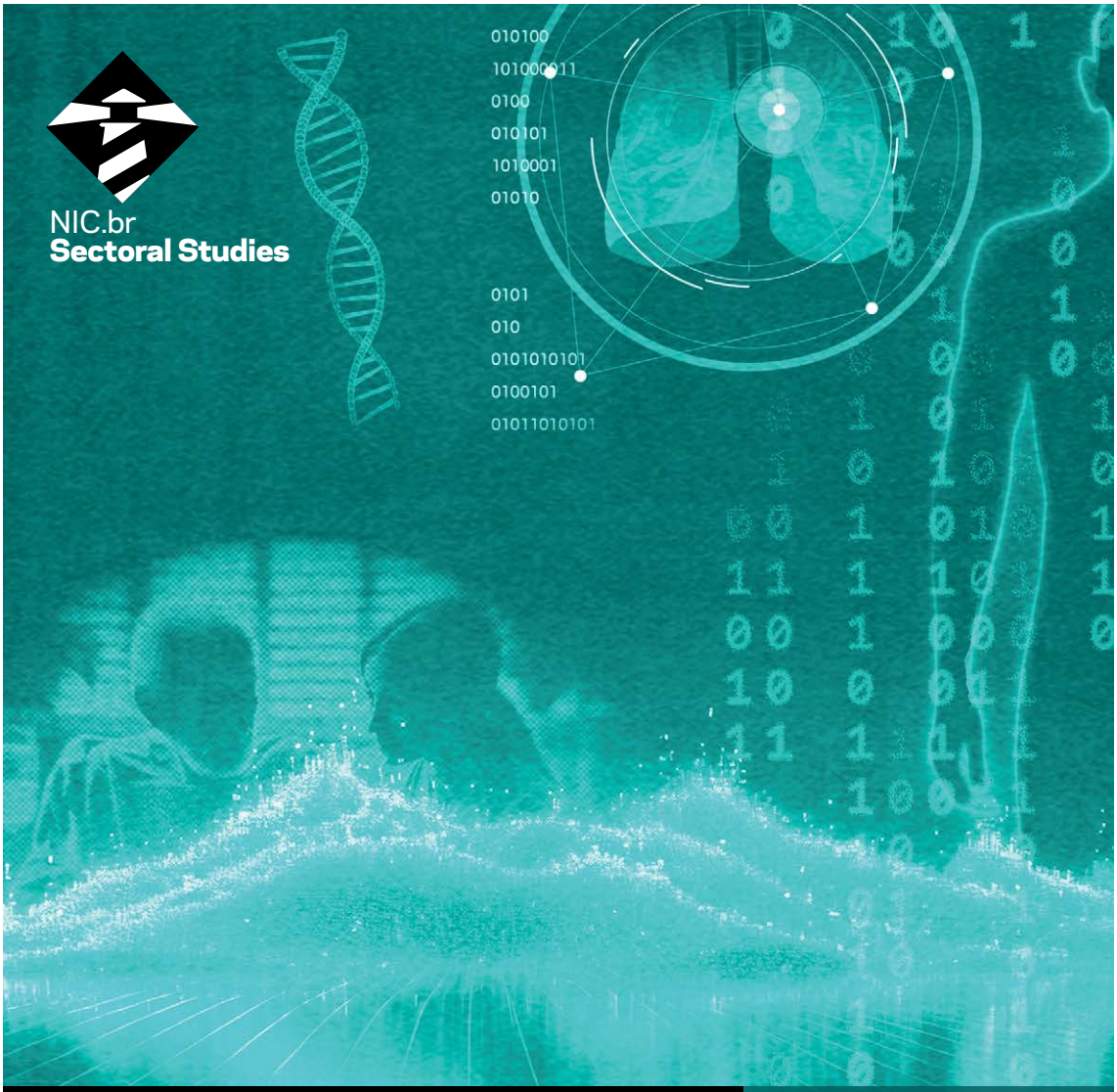


NIC.br
Sectoral Studies

ARTIFICIAL INTELLIGENCE IN HEALTHCARE:

potential, risks, and perspectives for Brazil

nic.br
Brazilian Network Information Center

**Brazilian Network Information Center –
NIC.br**

NIC.br
Sectoral Studies

ARTIFICIAL INTELLIGENCE IN HEALTHCARE:
potential, risks, and perspectives for Brazil

Brazilian Internet Steering Committee – CGI.br
São Paulo 2024

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Artificial intelligence in healthcare [livro eletrônico] : potential, risks, and perspectives for Brazil / Núcleo de Informação e Coordenação do Ponto BR (NIC.br) ; tradução Ana Zuleika Pinheiro Machado. -- São Paulo : Comitê Gestor da Internet no Brasil, 2024.
PDF

Título original: Inteligência artificial na saúde: potencialidades, riscos e perspectivas para o Brasil
Vários colaboradores.
Bibliografia.
ISBN 978-65-85417-62-4

1. Inteligência artificial 2. Pesquisa qualitativa 3. Saúde - Brasil 4. Saúde - Pesquisa I. Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

24-228786 CDD-610.3

Índices para catálogo sistemático:

1. Ciências da saúde : Pesquisa 610.3
Eliane de Freitas Leite - Bibliotecária - CRB 8/8415

Brazilian Internet Steering Committee – CGI.br

COORDINATOR
Renata Vicentini Mielli

COUNSELORS
Artur Coimbra de Oliveira
Beatriz Costa Barbosa
Bianca Kremer
Cláudio Furtado
Cristiano Reis Lobato Flôres
Débora Peres Menezes
Demi Getschko
Henrique Faulhaber Barbosa
Hermano Barros Tercius
José Roberto de Moraes Rêgo Paiva Fernandes Júnior
Lisandro Zambenedetti Granville
Luanna Sant'Anna Roncoratti
Luiz Felipe Gondin Ramos
Marcelo Fornazin
Marcos Adolfo Ribeiro Ferrari
Nivaldo Cleto
Pedro Helena Pontual Machado
Percival Henriques de Souza Neto
Rafael de Almeida Evangelista
Rodolfo da Silva Avelino

EXECUTIVE SECRETARY
Hartmut Richard Glaser

CONTENTS

ACKNOLEDGEMENTS

The Brazilian Network Information Center (NIC.br), through the Regional Center for Studies on the Development of the Information Society (Cetic.br), thanks all the professionals involved in the preparation of this publication. In particular, we thank Heimar de Fátima Marin (International Medical Informatics Association [IMIA]), the World Health Organization (WHO) and the International Telecommunication Union (ITU), Fernando Aith and Daniel A. Dourado (University of São Paulo [USP]), Monise Picanço and Priscila Vieira (Brazilian Center for Analysis and Planning [CEBRAP]), and Glauco Arbix and João Paulo Cândia Veiga (USP) for their contributions.

PRESENTATION

The Internet has become an essential infrastructure for developing new digital applications and for producing and disseminating data on an unprecedented scale. Applications based on Artificial Intelligence (AI) have benefited from components and the vast volume of information available on the network. Currently, AI permeates various areas of social life. In the health sector, research, and discussions about this technology are intensifying and becoming increasingly urgent. Thanks to machine learning (ML) techniques, AI has been able to promote significant advances in several sectors: from healthcare, including the prevention of pandemics, the development of new medications, and the optimization of healthcare facility management, to supporting decision-making by doctors, nurses, and other professionals in the field. In each of these sectors, numerous AI systems are in use or in advanced stages of development and testing, to assist human agents in overcoming both long-lasting challenges, such as making accurate diagnoses, and emerging challenges, such as rising healthcare costs and an aging population.

There is a clear promise that AI will transform healthcare, but its careful implementation will be crucial to maximizing the benefits and minimizing the risks of its adoption. Alongside the growth of initiatives to make AI an ally in improving healthcare, discussions about the ethical risks involved are also increasing. Concerns include the protection of personal data, the adequate training of healthcare professionals in the use of AI systems, and the fear that the potential benefits of the technology, such as personalized treatments, will only be accessible to the most privileged sections of society. In this context, regulatory debates are multiplying, guided by the need to establish guidelines that minimize the AI risks in health, without discouraging its development and experimentation.

For over 20 years, the Brazilian Network Information Center (NIC.br) has been working in collaboration with different stakeholders to promote an open and interoperable Internet, contributing to making a secure, inclusive, and high-quality network. Additionally, NIC.br has developed effective network security management mechanisms and offers a diversified portfolio of products and services aimed at continuously improving the Internet. In this context, NIC.br,

through the Regional Center for Studies on the Development of the Information Society (Cetic.br), has been producing data and analysis for a broad understanding of the effects of the adoption of digital technologies in Brazil, including AI. In this way, NIC.br plays an active role both in mapping the landscape of AI adoption in healthcare and in producing evidence to help decision-makers navigate the sector. NIC.br's initiatives to this end also include the Brazilian Artificial Intelligence Observatory (OBIA), serving as a focal point for monitoring and analyzing the evolution and impact of AI in Brazil. OBIA is the result of a strategic initiative within the context of the Brazilian Artificial Intelligence Strategy (EBIA) and the Brazilian Artificial Intelligence Plan (Plano Brasileiro de Inteligência Artificial [PBIA]), and its mission is to consolidate and disseminate knowledge about the impacts of AI on society by collecting and analyzing data on its adoption and use.

The indicators produced by Cetic.br stand out among the activities carried out by NIC.br, as they highlight the positive advances brought about by the expansion of the Internet and digital technologies in Brazil and point out the challenges that still need to be overcome for the population to benefit from the opportunities in a meaningful way. Among the indicators produced by Cetic.br|NIC.br are those of the ICT in Health survey, which, in its most recent editions, began to disclose data on the adoption of AI in the Brazilian healthcare sector.

The data released by Cetic.br|NIC.br are based on a multi-sectoral approach, from the planning of the methodology to the construction of data collection instruments, with the collaboration of experts from different areas. Disseminating this data to society helps to draw up policies and initiatives to improve both the technical and content layers of the Internet, in addition to promoting the expansion of tools available to the population and ensuring rights and critical, responsible, safe, and productive access to the Internet.

This publication is another effort by Cetic.br|NIC.br to understand the AI scenario in healthcare, and how it can serve as an instrument for society. It brings together a careful mapping of the main debates on AI and healthcare in the literature, highlighting issues related to transparency and explainability, as well as the perspectives of researchers,

healthcare facility managers, public sector and market representatives, and healthcare professionals on the challenges, risks, and opportunities posed by the technology. It also presents a series of regulatory considerations published by the World Health Organization (WHO) and the International Telecommunications Union (ITU). The report also includes notes for the development of public policies related to AI in the Brazilian healthcare sector.

Enjoy the reading!

Demi Getschko
Brazilian Network Information Center - NIC.br

PROLOGUE

Artificial Intelligence and health

Heimar de Fatima Marin[1]

1 Full professor (retired) at the Federal University of São Paulo (Unifesp), scientific coordinator of the ICT in Health survey at the Regional Center for Studies on the Development of the Information Society (Cetic.br) of the Brazilian Network Information Center (NIC.br), editor-in-chief of the International Journal of Medical Informatics (IJMI), and president of the International Academy of Health Sciences Informatics (IAHSI) of the International Medical Informatics Association (IMIA).

The term "Artificial Intelligence" (AI), coined over 65 years ago, has increasingly been used across all areas of human activity, including healthcare. The frequent use of this term reflects the expansion of studies, showing the advancements and applications in what is now called digital health, a more recent term employed by the World Health Organization (WHO) to replace and complement the concept of e-Health. Thus, the term digital health is used to emphasize the importance of consumer engagement in a healthcare system, broadening access to available resources.

The term AI has been widely used since the end of 1956 and saw its peak from the 1960s to the 1990s, with numerous studies developed and published in scientific journals in the fields of health and computing. After this period, there was a noticeable decline in the use of AI techniques in healthcare, suggesting that the main difficulty was integrating these decision-support systems into the electronic patient records, which, though still in the early stages, already contained important clinical data to support diagnosis, such as vital signs, symptoms, and major patient complaints seeking care. Thus, to use a knowledge-based support system or expert system, the professional needed to interrupt their current activity (such as taking the patient's history or performing a physical exam) to consult the system, which often operated independently as a standalone. After obtaining the results presented by the system, the professional would decide whether to incorporate these results into the record and follow the recommended diagnosis or treatment. Hence, these were typically isolated systems that relied on human interaction for consultation; upon accepting the recommendation, the professional assumed responsibility for the result presented by the system.

During this period, I began my studies in AI in healthcare, starting with my master's dissertation (1991) and doctoral thesis (1994), when AI was understood as a branch of computer science that employed computer programs capable of performing tasks normally associated with intelligent human behavior. From around 1985 to the early 1990s, some applications also stood out, such as the development of clinical decision support systems or expert systems, and the so-called

knowledge-based systems, which were already present in human behavior and decision-making. Specifically, these systems used reasoning to infer conclusions from stored facts. Thus, nothing that was developed could surpass the human capacity for thought, as everything they contained was the result of human cognitive ability.

In terms of basic architecture, these systems were composed of a knowledge base, an inference engine, and a user interface. The knowledge base contained rules, facts, concepts, and definitions of what was known about a particular topic at the time of its creation. The inference engine employed search strategies within this knowledge base to help solve problems, assist in diagnosis, recommend treatments, and communicate its findings to the user through the interface.

It's easy to see that the greatest challenge lay in building the knowledge base, since the more updated it was, the better the system would perform. Consequently, various methodologies were developed in an attempt to emulate human thinking, including algorithms, frames, semantic networks, neural networks, and even hybrid methods that utilized object-oriented programming techniques with production rules. Examples of these techniques include backward chaining, forward chaining, if-then rules, decision trees, pattern recognition, Bayesian theorem-based systems, and fuzzy logic, to name a few of those most commonly used by researchers at the time. Many studies employed backward chaining rules, where knowledge was represented by condition-action pairs. In this approach, a condition is an expression that must be true for the rule to be applied, and the action is a list of commands executed if the condition is met. The scientific principle of hypothesis refutation guided the process.

As a result, several systems developed from the early 1970s to the 1990s became global references and greatly contributed to advancements in implementations. Notable examples include MYCIN (antibiotic recommendation), INTERNIST-1 (used for teaching internal medicine, which later evolved into the Quick Medical Reference [QMR]), Leeds (based on Bayesian theorem, assisting in diagnosing acute abdominal pain), and HELP Systems (supporting hospital functions in clinical data management).

Among all these pioneering systems, it is worth noting the significant influence of the MYCIN system, which was powerful due to its representation and reasoning approach. Rule-based systems in many non-medical domains were developed in the years following MYCIN's introduction. However, despite their ease of understanding, these systems faced limitations when applied to large volumes of data.

Following the aforementioned period of decline from 2000 to 2018, marked by a significant decrease in scientific publications in health, a new peak began to emerge from 2018 onward, due to the enormous advancements in storage and processing capabilities for massive databases and deep learning techniques. With the exponential growth of knowledge and the increased adoption of digital technologies in healthcare, rule-based systems began to pose challenges in both their development and maintenance. Although the initial applications of machine learning techniques started in the 1960s, they have now gained significant prominence in all AI developments and have almost become synonymous with AI. The topics explored at that time remain very similar to current research focuses, but there is now a greater emphasis on diagnosis, treatment, and survival studies for all types of cancer, mental health, and neurological diseases.

This technological evolution has been driven by the massive explosion of health data, enabling the creation of advanced machine learning models known as deep learning. The availability of large volumes of data (provided it is high-quality data) has increased the capacity to solve problems that are not clearly described and defined. However, it remains critical to specifically identify the problem that can be addressed by the available AI base and the algorithms to be employed that bring clinical significance.

Since 2018, as mentioned, there has been a tremendous rise in the use and development of applications using AI and its most advanced techniques. Currently, the term dominates most studies, research, and investigations into how such resources can be applied in healthcare and all segments of human activity. Moreover, the issue of integration is now better resolved with resources for syntactic interoperability (format and order of what is exchanged) and semantic (meaning of

what is exchanged between systems) interoperability. As a result, resistance to using intelligent systems has become more easily mitigated.

In the research field, many studies continue to focus exclusively on comparing predictive models and algorithms, testing which ones perform better on a given set of health data. Although these studies provide diagnostic inferences, treatment recommendations, or survival expectations for certain pathologies, the most critical aspect is often missing: The number of studies conducted and published that demonstrate impact assessment and clinical significance is still minimal. Studies frequently use a sufficiently large dataset that can be divided into parts, allowing for data cleaning, followed by one part used to develop the model and a smaller part for internal validation testing. However, studies that test the model or algorithm on external datasets for external validation remain rare in the literature.

In education and training for new generations, the gap is even wider. Few professional training programs include health informatics and digital health topics as an integral part of the skills required for their graduates. Topics such as AI, machine learning, and similar predictive models are even rarer, due to the critical mass deficit in Brazil. This is crucial because we need professionals capable of using these resources, as the challenge remains: To use AI ethically, avoiding biases that can increase discrimination in the formation and selection of databases for clinically meaningful studies, and assisting professionals in the increasingly complex healthcare environment.

It is essential to translate the promises of AI into transforming healthcare models and programs implemented by national and international health systems, using its full potential. In the real world, these resources must be integrated into citizens' health records. It is crucial to ask, "Who benefits from the application of AI?" and "What are the consequences of its application?" Furthermore, it is always necessary to remember that human intelligence remains in control of the process.

Motivated by the topic and the scenario of new possibilities for use, this publication offers a relevant reflection on the stage of AI adoption in healthcare in Brazil, addressing the exploration of advances and challenges documented in the literature, with

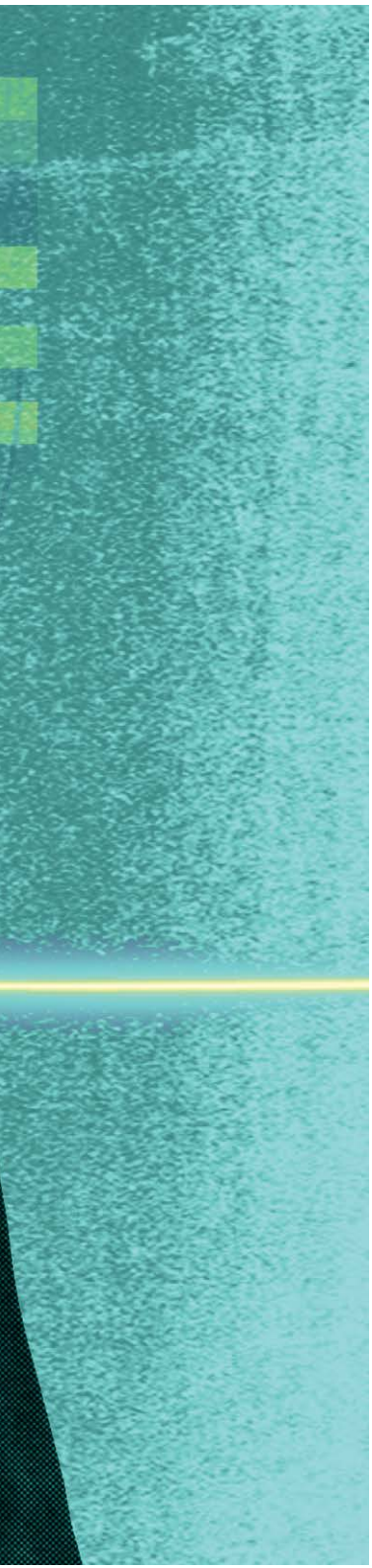
specific guidelines for the application of AI in the Brazilian context. It also highlights the most important and sensitive issue of the moment: The need for proper regulation to guarantee safety, efficacy, and ethics of AI usage in healthcare.

Part 1
ARTICLES

Artificial Intelligence in healthcare: An overview of the literature and guidelines for Brazil

Rodrigo Brandão[1]

1 Ph.D. candidate in Sociology at the University of São Paulo (USP), with a master's degree in Political Science and a bachelor's degree in Social Science at USP. He is a researcher at the Coordination of Qualitative Methods and Sectoral Studies at the Regional Center for Studies on the Development of the Information Society (Cetic.br) of the Brazilian Network Information Center (NIC.br).



Emerging information and communication technologies (ICT) and the possibilities arising from automation can bring great benefits to the health sector in its various stages: Prevention, diagnosis, the care, and treatment of individuals, as well as in the management of health systems and healthcare facilities, thereby optimizing the work of professionals, resources, operations, and established routines. Through the use of health data from multiple sources — such as electronic medical records, image acquisition, and storage, analysis of genomic profiles, and other physiological data — Artificial Intelligence (AI) also has great potential to help tackle challenges in the health sector, such as the continuous increase in costs, the lack of professionals, and the epidemiological and demographic changes underway, such as population aging.

The use of AI tools is also a promising approach for developing and implementing public policies on a large scale, and for leveraging and accelerating the production of scientific knowledge in the sector through research. Any optimism regarding technology, however, requires caution, given that the development and use of AI applications in healthcare are not without risks, such as the leakage of sensitive patient data, increased opacity in diagnoses, reduced accountability in medical decisions made using AI technologies that provide diagnostic, treatment, and prognostic suggestions, and even the widening of inequality in access to quality healthcare.

Faced with opportunities and risks as significant as these, it is worth asking: How can we take advantage of the potential benefits and mitigate the potential damage? In order to map out strategies that are capable of doing so, this chapter reviews the literature on the use of AI in healthcare.

The Section "Artificial Intelligence in health: A brief history and fundamental concepts," presents a brief history of AI, and highlights the uses of this technology in health over time. The polysemic meaning of the term "Artificial Intelligence" is also discussed, and it is pointed out that in health, as in other areas, it has been taken as a synonym for machine learning (ML) in recent years. In the Section "Current overview of AI uses in healthcare: Opportunities, challenges and risks," an updated panorama of the use of AI/ML applications in

healthcare is outlined, with reference to the administration of healthcare facilities and clinical practice to explore the opportunities and challenges for the advancement of AI, and the risks that this technology can bring.

The Section "Recommendations for the development and responsible use of AI: A review" addresses the recommendations of the World Health Organization (WHO) for the development and responsible use of AI in health in the world. In the Section "Guidelines for the development and responsible use of AI in healthcare in Brazil: The role of the DHS," the alignment between such recommendations and the Digital Health Strategy for Brazil 2020-2028 (DHS) (Ministry of Health, 2020) is evaluated and discusses how this state strategy can help structure the development of AI in the country. Finally, the concluding remarks are presented based on the key points analyzed in the previous sections (Section "Conclusion").

ARTIFICIAL INTELLIGENCE IN HEALTH: A BRIEF HISTORY AND FUNDAMENTAL CONCEPTS

In the 1950s, two events marked the birth of AI: mathematician Alan Turing released the article "Computing machinery and intelligence" in 1950, in which the Turing Test was mentioned for the first time; in 1956, during the Dartmouth Summer Research Project on Artificial Intelligence workshop, computer scientist John McCarthy and his colleagues coined the term "Artificial Intelligence" to describe "the science and engineering of making intelligent machines" (McCarthy, 1956 as cited in Berryhill et al., 2019, p. 11). AI was thus consolidated as a new area of research in engineering and computer science.

The definitions of AI have multiplied since then and there is no standard, globally accepted definition (Miailhe & Hodes, 2017). Aware of this, Krafft et al. (2020) sought to map how different audiences understand AI, based on two groups: Researchers in technical areas of AI and policymakers. Based on a comparison of the definitions used by each of them in different documents, such as academic articles and regulatory texts, they found that the former tend to use definitions that emphasize the technical functionalities of AI systems, while the definitions used by policymakers tend to be based on comparisons between AI applications and human ways of thinking and behaving.

Krafft et al. (2020) also observed that the definitions used by researchers in technical areas include AI applications that are already in use, while the definitions given by policymakers are more aligned with technologies that could hypothetically be created. Using two common terms in AI discussions, the definitions used by the first group align with the idea of Artificial "Narrow" Intelligence (ANI), while those of the second group express the concept of Artificial "General" Intelligence (AGI). ANI — often called weak AI — operates only in the scenarios in which it has been programmed to do so, performing specific tasks; AGI — known as strong AI —, on the other hand, can perform intellectual activities, such as abstracting and generalizing (Miailhe & Hodes, 2017).

Krafft et al. (2020) concluded that among the different definitions for AI they found, the one offered by the Organization for Economic Cooperation and Development (OECD) is the one that best fits between the views of these two groups:

> [...] An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. (OECD, 2024 as cited in Krafft et al., 2020, p. 6)[2]

It would be limiting, however, to say that the plethora of concepts in the field of AI is only due to differences in understanding between stakeholders in different fields. To prepare their textbook on AI, Russell and Norvig (2016 as cited in Krafft et al., 2020) analyzed computer science textbooks on the subject that were published between 1978 and 1993 and found that researchers in the field were interested in building four types of system that: (a) think like humans; (b) act like humans; (c)

2 The WHO (2021) adopted this definition in its report *Ethics and governance of Artificial Intelligence for health - WHO guidance*. However, it should be noted that in 2023, the OECD updated this definition to read: "AI system: An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." The new definition is available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

think rationally; and (d) act rationally. Sweeney (2003 as cited in Krafft et al., 2020), for her part, analyzed the 996 papers cited in the textbook prepared by Russell and Norvig (2016) and concluded that in 987 of them the researchers' attention — and therefore the definitions they used — were focused on developing systems that think and act rationally, or as Krafft et al. (2020) note, that think and act "ideally."

Even though the definitions for the central term ("Artificial Intelligence") revolve around a common axis ("rational/ideal reasoning and behavior") in the large area of AI, technicians from the area were unable to summarize them in a single, completely consensual formulation. Like any field of scientific knowledge, AI has subdivisions and the different definitions given to the central term are an expected reflection of this fragmentation. Among the subdivisions of the broad field of AI are areas such as ML, knowledge representation and reasoning, multi-agent systems, and others.

Among all these areas, academic and commercial research has been concentrating on the field of ML since the early 2000s (Arbix, 2020). Prior to this, AI went through a period known as the "AI winter," when funding and social and academic interest in the field became significantly scarce, which "froze" its development (Daugherty & Wilson, 2018; Kaul et al., 2020). There is no consensus on the exact start and end dates of this phase, but it is generally understood to have lasted from the mid-1970s to the early 2000s.

As discussed further, the relationship between AI and health has always been close, even during periods of slow development; medicine has been considered a promising area for application of the technology since the early days of AI (Yu et al., 2018). According to Kaul et al. (2020), the first phase of AI extended from the 1950s to the mid-1970s, a time when the focus was on developing systems capable of making inferences or decisions previously made exclusively by humans. In the medical field, the main advance in this period was the digitization of data, with the creation of medical records systems and clinical informatics databases, with the development in 1960 of the Medical Literature Analysis and Retrieval System and PubMed by the American National Library of Medicine.

In subsequent years, the results obtained by engineering and computer science in AI fell short of expectations. This frustration turned into what became known as the "first AI winter," which lasted until the early 1980s. Interest in the area regained momentum when some computer science researchers managed to equip computer systems with if-then reasoning logic. Advances like this were not enough, however, to re-establish the strength of the area, as the costs involved were high, especially for maintaining specialized databases. Another event taking place at the same time also attracted the attention of funders and researchers, namely the development of the desktop (Daugherty & Wilson, 2018). Together these factors gave rise to the "second AI winter," which ended around the 2000s.

AI has seen important developments over the years in the medical field, however, from the creation of academic organizations to strengthen the relationship between AI and medicine, to the development of medical systems centered on the technology in question. Among the latter were decision support systems such as MYCIN and DXplain, which were developed in the early 1970s and late 1980s respectively.[3] The development of AI continued in subsequent years, but it was from 2010 onwards that it gained traction, due to three factors: (a) personal databases were growing rapidly, thanks to app-centric economic and social dynamics; (b) the technological infrastructure was developed to store this information; and (c) the field of statistics was ready to analyze it (Arbix, 2020).

Like other areas of AI, the ML area is dedicated to building algorithms, which can be defined as "[...] encoded procedures for transforming input data into a desired output, based on specified calculations" (Gillespie, 2014, p. 167). In the case of ML, researchers develop procedures for identifying patterns in large volumes of data and work to ensure that computer systems learn these procedures. Having learned the pattern of a cancer image, for example, the technology is now able to estimate the probability that a new image will match the

3 Developed by Stanford professor, Edward Shortliffe, MYCIN stood out among decision support systems due to its contributions to both medicine and computer science (van Melle, 1978).

learned pattern. Frame 1 shows the three most common types of ML algorithms.

FRAME 1 - MAIN TYPES OF ML ALGORITHMS

Supervised ML: A type of ML task that aims at predicting the desired output (such as the presence or absence of diabetic retinopathy) on the basis of the input data (such as fundus photographs). Supervised ML methods work by identifying the input-output correlation in the "training" phase and by using the identified correlation to predict the correct output of new cases.

Unsupervised ML: A type of ML task that aims at infering underlying patterns in unlabeled data. For example, it can find sub-clusters of the original data, identify outliers in the data, or produce low-dimensional representations of the data.

Deep learning (DL): A subfield of the larger discipline of ML. DL employs artificial neural networks with many layers to identify patterns in data.

SOURCE: ADAPTED FROM YU ET AL. (2018).

According to Topol (2019), DL algorithms are distinguished by their self-learning capabilities. In general terms, when exposed to a set of annotated images, such as cancer and non-cancer, supervised learning algorithms learn to recognize common characteristics of these images, for example, the volume of a nodule. Composed of layers reminiscent of human brain neural networks, deep learning algorithms can evaluate numerous aspects of analyzed images, without the developers fully understanding which specific aspects the algorithms considered when generating the results. For this reason, it is often said that many DL algorithms resemble "black boxes," as they do not allow us to know precisely how inputs are converted into outputs. In the health sector, this component poses important challenges in terms of the ethical, transparent, and explainable decisions that need to be taken.

Due to the logic of the way they function, DL algorithms have a superior capability to that of their peers when it comes to recognizing patterns in large volumes of data, thus demonstrating that the new generation of AI systems is quite distinct from previous generations, which relied on expert curation of health knowledge and robust decision rules (Yu et al., 2018). This significant shift may gain momentum in the medium term, given that private investments in AI in health have been

consistent. The *Artificial Intelligence index report 2023* (Stanford University Human-Centered Artificial Intelligence [HAI], 2023) notes that medical research was the primary destination for these resources in 2022: US$ 6.1 billion were allocated to it, compared to US$ 5.9 billion for research into administration, data processing and cloud computing, and US$ 5.5 billion for developments related to fintechs. Given the close relationship between AI and health, it is important to understand the opportunities, challenges, and risks involved. The Section "Current overview of AI uses in healthcare: Opportunities, challenges, and risks" is dedicated to mapping this.

CURRENT OVERVIEW OF AI USES IN HEALTHCARE: OPPORTUNITIES, CHALLENGES AND RISKS

Initially, information on the relationship between AI and health was sought in specific scientific journals, such as *Nature Biomedical Engineering* and *Nature Medicine*, and gray literature was used, such as government reports and analyses produced by consulting firms. For example, the *Technology assessment - Artificial Intelligence in health care: Benefits and challenges of machine learning technologies for medical diagnostics,*[4] published in 2022 by the United States Government Accountability Office and the US National Academy of Medicine, and *Artificial Intelligence in healthcare - Application, risks, and ethical and societal impacts,*[5] published by the European Parliament also in 2022, were analyzed.

At the end of the first round of the literature review, four themes were identified for in-depth investigation: (a) the use of AI techniques in biomedical research, especially for the development of new drugs; (b) the use of AI/ML systems in clinical practice to improve diagnoses, prognoses and treatments; (c) the use of AI/ML applications in the administration of healthcare facilities; and (d) the ethical and regulatory challenges associated with AI on these three fronts. Academic documents relating to the last three topics were searched for between May 10 and 15, 2023 in the Scopus database.[6] For each of the topics,

4 Find out more: https://www.gao.gov/products/gao-22-104629
5 Find out more: https://www.europarl.europa.eu/thinktank/pt/document/EPRS_STU(2022)729512
6 Find out more: https://www.scopus.com/home.uri

a specific search was carried out consisting of four stages.

In all three cases, the following keywords and Boolean operators were used in the field "Article title, Abstract, Keywords": "artificial intelligence" OR "machine learning" OR "deep learning" AND "health" OR "public health" OR "healthcare" OR "health care". These keywords were also combined in the first stage of the search, using the Boolean operator "AND" with three different sets of keywords: (a) "clinical practice" OR "diagnostic decision support" OR "diagnostic decision" OR "clinical decision support" OR "clinical decision"; (b) "healthcare management" OR "healthcare management system"; e (c) "ethics" OR "regulation".

Secondly, three filters were applied to the results found in the previous stage: (a) "year" — the period selected was "2019-2023"; (b) "type of document" — in the case of "clinical practice" only reviews were selected, while in the fields of "healthcare facilities management" and "ethics and regulation", both reviews and articles were selected; and (c) "language" — the languages selected were English, Spanish and Portuguese.

The third step was to order the results of the second stage in two different ways: (a) from the most to least cited, and (b) from the most to least relevant. In the following stage the titles and abstracts of the top 20 documents — ranked either by the number of citations or by their relevance — were read. Some of them were common to both lists. A total of 113 documents related to clinical practice, 94 related to the administration of healthcare facilities, and 219 documents on ethical and regulatory challenges were reviewed.

Three exclusion criteria were adopted to allow the selection of materials that together provide an overview of the benefits, challenges, and risks of AI in healthcare: (a) documents on the testing of AI models; (b) documents that address the benefits, challenges and risks of AI in a speculative manner; and (c) documents that focused on a single country (with the exception of a few documents on Brazil), or on a single disease (except for two texts discussing the role of AI in management challenges related to the COVID-19 pandemic). The use of these criteria culminated in the selection of 21 documents on clinical practice; five on the administration of healthcare facilities; and 32

documents on ethical and regulatory issues (Appendix 1[7]).

The Subsections "The use of AI/ML systems in clinical practice" and "The use of AI/ML systems in the administration of healthcare facilities" present a summary of the reading conducted in the two rounds of literature review on clinical practice and administering healthcare facilities, respectively. References to ethical and regulatory challenges are presented throughout these two subsections and the Section "Recommendations for the responsible development and use of AI: A review," along with WHO (2021) recommendations for the responsible development and use of AI.

THE USE OF AI/ML SYSTEMS IN CLINICAL PRACTICE

In general, the documents reviewed present the results of studies in which AI/ML algorithms performed well in interpreting clinical symptoms in different medical areas. Topol (2019) cites examples of this in radiology, pathology, dermatology, ophthalmology, cardiology, gastroenterology, and mental health. The author notes, however, that "validation of the performance of an algorithm in terms of its accuracy is not equivalent to demonstrating clinical efficacy" (Topol, 2019, p. 45).

However, at least in the United States, AI/ML algorithms may be reaching clinical practice without proper scrutiny from the academic community and/or government authorities (Topol, 2019), because many studies on the effectiveness of these algorithms are not published in peer-reviewed journals. The Food and Drug Administration (FDA) has also facilitated the approval process for medical algorithms in recent years, and government programs such as Medicare and Medicaid have started to reimburse organizations that adopt AI systems, while in other sectors, institutions interested in using the

7 Appendix 1 presents a table summarizing general information about the studies selected for reading. There are two observations regarding this list. The first is that the work by Pap & Oniga (2022) appeared in the search results for both clinical practice and ethics and regulation. After reading the summary, it turned out to be relevant to both areas; for this reason, Appendix 1 has 57 documents instead of 58. The second observation is that, after reading the full texts, some turned out to be more pertinent to the discussion of a topic other than the one that prompted the search. Such is the case with Hobensack et al. (2023), for example: Although the authors' work was identified during searches for documents related to the role of AI systems in clinical practice, it turned out to be more relevant to the discussion about opportunities and challenges linked to the administration of healthcare facilities. Despite this, the table lists the texts according to the three themes on which the Scopus searches were based.

technology must pay for it (Sahni et al., 2023).

Aware of the FDA's move, Benjamens et al. (2020) mapped the algorithms and medical devices that are based on AI/ML and approved by the US agency. The authors evaluated 7,390 official FDA announcements on products approved by the agency and found 64 algorithms and devices that could be based on AI/ML techniques. In 29 cases, Benjamens et al. (2020) noted that the official announcement contains the information that the approved product is based on AI/ML techniques, while in the other 35 cases, other sources of information (other than official announcements) indicate that the approved products are based on AI/ML techniques. The authors, therefore, included only the first set in their database. They point out that:

> The two main medical specialties with AI/ML-based medical innovations are Radiology and Cardiology, with 21 (72.4%) and 4 (13.8%) FDA-approved medical devices and algorithms respectively. The remaining medical devices and algorithms can be grouped as focusing on internal medicine/endocrinology, neurology, ophthalmology, emergency medicine, and oncology. The medical field of radiology is the trendsetter regarding FDA-approved medical devices and algorithms [...]. (Benjamens et al., 2020, p. 2)[8]

The work by Benjamens et al. (2020) also shows that the dissemination of AI/ML-based algorithms and devices in clinical practice has gained prominence in the scientific community. The authors noted that the FDA did not approve any products embedded with AI/ML in 2010 and 2011, while in 2012 there were two approvals. This number jumped to 22 in 2019 and reached 64 in 2020.[9] Below are 12 documents from

8 Few studies have been found that compare different medical areas, in order to identify in which of them AI/ML systems are most used. One of them is Tran et al. (2019), who identified the number of academic articles on AI techniques published for 25 diseases.

9 Benjamens et al. (2020) do not provide technical information on the algorithms and devices mapped, but Ngiam & Khor (2019) and Ahmed et al. (2020) partially fill this gap. The first mapped clinical trials in oncology focused on ML algorithms. Ahmed et al. (2020) mapped examples of ML algorithms used in different areas of health and identified which analysis method is used in each of them, such as DL, logistic regression and linear regression.

the literature review that provide insights into the challenges and risks associated with the integration of technology at healthcare frontlines.

The first is the study by Topol (2019), who analyzed 27 peer-reviewed publications and compared the performance of AI algorithms and physicians when it came to interpreting clinical data in different areas. The author's main conclusion is that "the field clearly is far from demonstrating very high and reproducible machine accuracy [...] for most medical scans and images in the real-world clinical environment" (Topol, 2019, p. 45). The author reached a similar conclusion when analyzing AI/ML algorithms aimed at predicting clinical outcomes.[10] He mapped 13 reports on ML and DL algorithms for predicting various outcomes, such as suicides and mortality rates after chemotherapy treatment. When analyzing them he noted important technical challenges, such as the heterogeneity of the cohorts studied and the accuracy range. For these reasons, the author believes that "it is not yet known how well AI can predict key outcomes in the healthcare setting, and this will not be determined until there is robust validation in prospective, real-world clinical environments, with rigorous statistical methodology and analysis" (Topol, 2019, p. 49).

At least in the short term, Topol's (2019) conclusions may contrast with the potential of AI/ML to improve clinical work. Therefore, it remains to be seen how the stakeholders involved in the adoption of technology at the healthcare frontline view the opportunities and challenges associated with it since their perceptions can impact both the development and use of AI/ML systems in healthcare.

The works by Yang et al. (2021), Hogg et al. (2023), and Aquino et al. (2023) help to fill this gap. The first two are literature reviews. Yang et al. (2021) sought to identify the perception of different actors regarding the impact of AI on radiology. The authors' main conclusion is that clinicians, surgeons, students, and patients are optimistic about the technology, with the caveat that realizing its potential depends on education and training for the professionals involved.

10 In general, the documents reviewed focus on interpreting, rather than predicting clinical results. Topol (2019) is one of the few authors to address both topics.

Hogg et al. (2023) investigated the factors that influence the adoption of AI systems in clinical practice. They focused on five groups of stakeholders: Developers; healthcare professionals; healthcare leaders and managers; patients, caregivers, and the general public; and regulators and policymakers. The authors' conclusions on healthcare professionals are particularly noteworthy.[11] The authors observed that three factors have an impact on the implementation of the technology in question for this audience: (a) a sense of their ability to understand AI systems; (b) a perception of how technology can change the relationship between them and their patients; and (c) the possibility of aligning the changes brought about by AI systems with current care behaviors.

Finally, Aquino et al. (2023) sought to understand whether the use of AI technologies in healthcare can harm the technical capacity of professionals in the field. To this end, they conducted 72 semi-structured interviews with different professionals involved in the development, use and regulation of AI systems. The authors found two contrasting views. "The utopian view was that AI could enhance existing clinical skills and systems, while the dystopian view was that AI would lead to replacement of tasks or roles by automation" (Aquino et al., 2023, p. 5).

Another eight selected papers discuss the use of AI in clinical practice with a reference to clinical decision support systems (CDSS), thus complementing the perception studies by Yang et al. (2021), Hogg et al. (2023), and Aquino et al. (2023). They have two common characteristics. The first is their focus on existing CDSS today, which — because they are based on Artificial "Narrow" Intelligence — provide suggestions for courses of action to their users instead of making autonomous decisions. The second characteristic common to these works is their focus on CDSS that help frontline healthcare professionals. In other words, they do not delve into clinical decisions of an administrative nature, such as prioritizing patients, nor into CDSS whose primary user is the patient themself, such as mobile healthcare apps for monitoring and controlling diabetes mellitus (El-Sappagh et al. 2019). In summary, the CDSS

11 Asan et al. (2020) offer additional reflections on the challenges encountered by clinicians when using AI/ML systems.

discussed correspond to the assistance algorithms in Frame 2.

FRAME 2 - LEVELS OF AUTOMATION OF AI/ML ALGORITHMS IN HEALTHCARE

	ASSISTIVE AI ALGORITHMS		AUTONOMOUS AI ALGORITHMS		
	LEVEL 1 DATA PRESENTATION	LEVEL 2 CLINICAL DECISION-SUPPORT	LEVEL 3 CONDITIONAL AUTOMATION	LEVEL 4 HIGH AUTOMATION	LEVEL 5 FULL AUTOMATION
Event monitoring	AI	AI	AI	AI	AI
Response execution	Clinician	Clinician and AI	AI	AI	AI
Fallback	Not applicable	Clinician	AI, with a backup clinician available at AI request	AI	AI
Domain, system, and population specificity	Low	Low	Low	Low	High
Liability	Clinician	Clinician	Case dependent	AI developer	AI developer
Example	AI analyses mammogram and highlights high-risk regions	AI analyses mammogram and provides risk score that is interpreted by clinician	AI analyses mammogram and makes recommendation for biopsy, with a clinician always available as backup	AI analyses mammogram and makes biopsy recommendation, without a clinician available as backup	Same as level 4, but intended for use in all populations and systems

SOURCE: BITTERMAN ET AL. (2020, AS CITED IN ADLER-MILSTEIN ET AL., 2022).

The eight studies can be divided into two groups. The first of these comprises three documents that analyze the use of CDSS in specific medical areas. Commissioned by the *British Medical Bulletin*, the work by Bishara et al. (2022) analyzed 13 examples of ML algorithms used in intensive care environments. Based on this framework, the authors discuss seven obstacles to implementing technology in the reality they investigated: (a) safety/accountability/liability; (b) interpreting ML findings to

patients; (c) privacy/anonymity; (d) ethics/fairness/equity; (e) data access and availability for ML and its generalizability; (f) regulatory approval; and (g) economic considerations.

Moazemi et al. (2023) carried out a systematic literature review of CDSS in cardiovascular intensive care units. The authors evaluated 21 academic articles and concluded that many of the results described had not been validated by external databases, thus revealing weaknesses when it came to generalizing these results. According to the authors, this challenge is becoming more critical as databases increasingly prioritize confidentiality. Moazemi et al. (2023) also concluded that the interpretability of the results generated by AI/ML systems is essential for medical teams to trust them.

The work by Du et al. (2023), which is also a systematic literature review, analyzed the use of CDSS based on ML involving various aspects of gestational care. The authors observed that "black box" algorithms are increasingly being used, and that clinicians have positive opinions of the CDSS they utilize, despite the common lack of explainability associated with these algorithms.

The other five papers discuss challenges associated with CDSS without focusing on specific medical areas. The first of these is the work by Sutton et al. (2020), which discusses seven pairs of opportunities and challenges for CDSS. The authors list topics such as patient safety versus healthcare professionals' fatigue due to alerts generated by CDSS, and the increased capacity for interpreting tests versus the need for the interoperability of data between systems from different institutions. Sutton et al. (2020) do not analyze AI/ML-based CDSS in depth: They only state that they can be important allies when defining diagnoses.

Commissioned by the United States Government Accountability Office and the United States National Academy of Medicine, the study by Adler-Milstein et al. (2022) discusses eight factors that can impact the adoption of CDSS for defining diagnoses. These topics include financial incentives for adopting these systems, such as government reimbursement; the necessary infrastructure for their proper functioning, including data interoperability and the processing power of the computers used; the quality of the diagnostic systems'

interface to prevent user fatigue among professionals; and the confidence physicians have in the results generated, which, according to the authors, depends on the explainability of these results.

Amann et al. (2020) focus on the explainability of the results suggested by the CDSS. The authors tried to understand the relevance of this element from four different perspectives: Technical, legal, medical, and patient. Based on a literature review, they concluded that the use of CDSS based on opaque "black box" algorithms can have at least two harmful consequences: (a) relegating the patient to a position of observer in the medical decision-making process; and (b) compelling doctors to adhere strictly to the results generated by the systems under discussion, in order to avoid being sanctioned legally and by their peers.

Albahri et al. (2023), in turn, evaluated various elements needed to build reliable AI algorithms in the health sector. One of the authors' conclusions is that there are few high-quality databases publicly available for this purpose, which indicates the impact that data quality has on the results produced by AI systems.

Finally, the fifth and last work is a systematic literature review conducted by Xu et al. (2023). Using 20 academic articles from 16 different journals as a reference, the authors investigated the interpretability of results generated by CDSS from the technological and medical perspectives.[12] Among their conclusions, they highlighted the mapping of analysis methods that ensure interpretability and found that the concept varies between doctors and patients. For the former, technical elements, such as biomarkers and the quality of the data used in CDSS, have a relevance not observed among patients, for whom interpretability is associated with promoting informed consent and facilitating their participation in treatment processes.

Without dwelling on CDSS, other works reviewed discuss the challenges of explainability. This group includes research

12 Following the majority of studies reviewed, Xu et al. (2023) do not propose a clear distinction between interpretability and explainability. The authors recognize that the results generated by a CDSS must be interpretable to facilitate explanation. For an in-depth discussion of the relationship between the two concepts, see Markus et al. (2021).

by Markus et al. (2021), Loh et al. (2022), and Liu et al. (2022). The first of these maps out and discusses different methods for ensuring the explainability of AI systems. Loh et al. (2022), for their part, carried out a systematic literature review based on 99 articles published in academic journals classified as Q1 in the SCImago Journal & Country Rank (SJR), from which they identified which AI explainability techniques are most used, without restricting themselves to specific diseases or databases. The work by Liu et al. (2022) is an opinion article in which the authors present a framework for auditing medical algorithms based on AI/ML.

As is evident from the reviewed texts, healthcare professionals must trust AI/ML systems in order to adopt them, and explainability is essential for this purpose. More precisely, the papers analyzed suggest that doctors, nurses, and other professionals working at the healthcare frontline need to be sure that they will not be misled by technology, which could cause harm to patients and expose healthcare professionals to lawsuits. They must also feel that they can demonstrate to their patients why they (healthcare professionals) have chosen certain diagnoses or courses of treatment over others. These two sets of concerns dialogue with the two elements that, according to Floridi et al. (2018), constitute explicability, namely, intelligibility ("How do AI systems work?") and accountability ("Who is responsible for how AI systems work?").[13]

The material consulted, however, does not really explore in any detail how the various stakeholders involved in AI/ML systems can collaborate to foster trust in the results that are generated. There are doubts among healthcare professionals, for example, about who would be primarily responsible for dealing with the challenges posed by "black box" algorithms: Themselves — through investments in education — or the developers of the technology (Hogg et al., 2023).

There are other unanswered questions. Topol (2019) problematizes, for example, the fact that physicians need to trust AI/ML algorithms in order to adopt them, while, at the same time, they prescribe medications whose action mechanism

13 The reviewed studies dedicate more attention to intelligibility than to accountability.

is unknown. Along these same lines, Markus et al. (2021) note that other elements may be essential for making AI algorithms reliable in the health field, such as publicizing the quality of the data used and extensive external validation of the results obtained. Considerations such as these suggest that the role of explainability — or, in the term used by Xu et al. (2023), types of interpretability — in building the trust of healthcare professionals in AI/ML systems requires new rounds of academic research.

Taken together, the works presented lead to two conclusions. The first is that analyzing the performance of AI/ML systems in clinical practice requires investigations that consider both technical factors, such as the quality of the databases, and human factors, such as confidence in the results generated. The second is that the challenges and risks of adopting these systems are diverse and of a socio-technical nature. Below is a list of the five challenges and risks identified as priorities in the selected material. Before discussing them, two observations are necessary.

The first consideration is that, in general, the works reviewed refer to ethical challenges and risks, since they can cause damage to the dignity of both healthcare professionals and their patients — whether in legal, moral, or physical terms. Despite the constant reference to the term "ethics," few authors explain which theoretical framework they are based on. The work by Amann et al. (2020) is one of the few exceptions: The authors point out that their considerations on the challenges and risks of AI in health are anchored in the four principles of bioethics: Beneficence, non-maleficence, autonomy, and justice .

Floridi et al. (2018) analyzed how these four principles appear in AI ethics documents from different areas other than just health. According to the authors, the first principle can be understood as the moral imperative — "do only good." In the documents they studied, beneficence is usually related to the advancement of the common good. "Non-maleficence" — understood as the moral imperative — "do no harm" — is usually associated with the care needed to protect privacy. As for "autonomy," Floridi et al. (2018) point to the need for human beings to have the power to choose in which situations they want to delegate decision-making powers to intelligent

systems and in which situations they want to revoke that decision. The principle of "justice" meanwhile, refers to the use of AI to promote prosperity and preserve solidarity. Finally, the authors point out that explicability is a principle that appears in a diffuse way among the documents they analyzed, revolving around ideas of intelligibility, transparency, and accountability.

The second observation about the challenges and risks identified is that the terms "challenges" and "risks" tend to be used interchangeably, since challenges can be seen as potential risks. However, to make the visualization of challenges and risks associated with the adoption of AI/ML systems in clinical practice clearer, the following emphasize the challenges by seeking to elucidate both the risks they entail and the ethical principles they engage with.

Challenge 1 - Preserving patients' privacy. The development and operation of AI/ML systems requires large volumes of data, so the first challenge is to ensure massive data collection that respects individual privacy. Gaps in this area can result in harm to patient autonomy (Bishara et al., 2022). In the name of beneficence, however, privacy must be reconciled with data collection that can improve individual and collective health. Once health data has been collected it needs to be protected against various dangers, such as leaks, unauthorized sharing by its holders and potentially discriminatory uses based on profiling techniques, such as charging higher health insurance premiums to patients whose personal data suggests likely adverse health conditions in the future (Bishara et al., 2022).

Challenge 2 - Ensuring the quality and representativeness of the data used. Authors such as Moazemi et al. (2023) point out the need to validate AI/ML algorithm results in health using databases that are different from those they were trained on, thereby increasing the chances of generalizing their findings. Albahri et al. (2023), for their part, identified a lack of high-quality public databases for this purpose. Finally, Challen et al. (2019) mapped the risks associated with problems in the data used to develop AI/ML systems (listed and described in Frame 3).

In this context, it is important to note that, although health data can be of different types, a significant portion

corresponds to electronic health records (EHR), which, according to Bishara et al. (2022), tend to be fragmented, suffer from poor formatting and missing information, including unstructured data, and may not always accurately reflect the clinical situation they refer to. For this reason, the data associated with EHR requires careful technical processing before being used in AI/ML systems.

The works by Topol (2019), Challen et al. (2019), and Schwalbe & Wahl (2020) suggest two other challenges that increase the risks in Frame 3. The first is the lack of consensus on how to report or even compare the accuracy of AI/ML systems. The second is the lack of tests in real clinical environments. Both are partly due to the fact that many experiments using AI/ML are not peer-reviewed.

In addition to being extensive and processed, the data to be used in the systems under analysis must be representative, especially in terms of race, gender, sexual orientation, and age. This is because it is socially undesirable, according to the principle of justice, for technology to be proficient in identifying, for instance, melanomas in people with lighter skin tones, but not in darker skin tones. Giovanola & Tiribelli (2023) list and discuss biased and discriminatory results of this nature; having as their reference thinkers such as John Rawls, the researchers also address the possibility of AI/ML systems accentuating socioeconomic inequalities in access to and the use of quality health services.

FRAME 3 - RISKS ASSOCIATED WITH THE VOLUME AND QUALITY OF DATA

ISSUE	SUMMARY	EXEMPLE
SHORT TERM		
Distributional shift	A mismatch between the data or environment the system is trained on and that used in operation, due to bias in the training set, change over time, or use of the system in a different population, may result in an erroneous "out of sample" prediction	The accuracy of a system which predicts impending acute kidney injury based on other health records data, became less accurate over time as disease patterns changed
Insensitivity to impact	A system makes predictions that fail to take into account the impact of false positive or false negative predictions within the clinical context of use	An unsafe diagnostic system is trained to be maximally accurate by correctly diagnosing benign lesions at the expense of occasionally missing malignancy
Black box decision making	A system's predictions are not open to inspection or interpretation and can only be judged as correct based on the final outcome	A X-Ray analysis AI system could be inaccurate in certain scenarios because of a problem with training data, but as a black box this is not possible to predict and will only become apparent after prolonged use
Unsafe failure mode	A system produces a prediction when it has no confidence in the prediction accuracy, or when it has insufficient information to make the prediction	An unsafe AI decision support system may predict a low risk of a disease when some relevant data is missing. Without any information about the prediction confidence, a clinician may not realise how untrustworthy this prediction is
MEDIUM TERM		
Reinforcement of outmoded practice	A system is trained on historical data which reinforces existing practice, and cannot adapt to new developments or sudden changes in policy	A drug is withdrawn due to safety concerns but the AI decision support system cannot adapt as it has no historical data on the alternative
Self-fulfilling prediction	Implementation of a system indirectly reinforces the outcome it is designed to detect	A system trained on outcome data, predicts that certain cancer patients have a poor prognosis. This results in them having palliative rather than curative treatment, reinforcing the learnt behaviour

LONG TERM		
Negative side effects	System learns to perform a narrow function that fails to take account of some wider context creating a dangerous unintended consequence	An autonomous ventilator derives a ventilation strategy that successfully maintains short term oxygenation at the expense of long-term lung damage
Reward hacking	A proxy for the intended goal is used as a "reward" and a continuously learning system finds an unexpected way to achieve the reward without fulfilling the intended goal	An autonomous heparin infusion finds a way to control activated partial thromboplastin time (aPTT) at the time of measurement without achieving long-term control
Unsafe exploration	An actively learning system begins to learn new strategies by testing boundary conditions in an unsafe way	A continuously learning autonomous heparin infusion starts using dangerously large bolus doses to achieve rapid aPTT control

SOURCE: ADAPTED FROM CHALLEN ET AL. (2019).

The WHO (2021) also draws attention to the representativeness of health data from low- and middle-income countries (LMIC). According to the institution, AI systems are mainly being used in the United States and the European Union, which could have a negative impact on the LMIC in two ways. The first is that AI/ML systems tend to be poorly exposed to health data from these countries, so the technology could fail to benefit them. The second consequence is that discussions about the development and responsible use of AI/ML systems in health tend to take into account the socio-legal mechanisms that exist in rich countries. Mechanisms in the LMIC can vary significantly, as noted by the WHO (2021). Moreover, many countries have a limited regulatory capacity in health. For these reasons, the institution states that it is still unknown which approaches LMIC will use to address the challenges and risks associated with AI, and it is also unclear how they will seek to prevent digital exclusion from hindering the potential of AI to improve health outcomes, something that is already significant in some of them.

Challenge 3 - Ensuring the (data) infrastructure necessary for the technology to function properly. Due to the

characteristics of the EHR, it is common for medical technologies to operate in very specific contexts, and AI/ML systems add layers of complexity to this reality. One of these is the need for AI/ML algorithms to be exposed to new quality data in order to be updated, which can lead to users finding themselves in a situation of technological dependence, being at the mercy of developers or specific data providers (lock-in) (Centre for the Fourth Industrial Revolution Brazil [C4IR], 2022). The proper functioning of some AI/ML algorithms may also require data from different systems, thus requiring the creation of interfaces that make them interoperable. This data, in turn, needs to be collected from some data-generating source, such as an electrocardiogram machine, stored either in a "data lake" or a "data warehouse," and processed.

Challenge 4 - Developing systems that are attentive to the user experience. The works by Sutton et al. (2020) and Adler-Milstein et al. (2022) point out that CDSS cannot generate an excessive number of alerts and that the information generated must be easily accessible. Another risk is the lack of intelligibility and interpretability, which can expose patients to physical risks, and expose doctors, nurses, and other healthcare workers to professional and legal risks.

Challenge 5 - Ensuring training for frontline healthcare professionals. Doctors, nurses, and other professionals at the healthcare frontline are not primarily responsible for solving the challenges indicated. If they are prepared to interact with AI/ML systems, however, they will have a better chance not only of identifying dangerous situations for themselves or their patients but also of using technology to improve their professional skills, thus benefiting the patients they serve. In other words, educating and training professionals at the healthcare frontline is essential for strengthening human supervision over technology, so that human experience and judgment are integrated into the functioning of AI systems. Known as human-in-the-loop, this formula is not sufficient to remedy the risks of AI/ML systems, since it can itself be flawed (Frame 4). Even so, it can help control the risks in Frame 3.

By way of conclusion, it should be emphasized that the existence of explainability in this literature review proved to be a crucial challenge if AI/ML systems are to be considered

transparent and, therefore, reliable. Explainability in the list of challenges breaks down into "Challenge 2 - Ensuring the quality and representativeness of the data used" and "Challenge 4 - Developing systems that are attentive to the user experience."

FRAME 4 - PROBLEMS RELATED TO THE HUMAN SUPERVISION OF AI SYSTEMS

AUTOMATION BIAS	Excessive reliance on algorithm-generated results can render human supervision and the role of the human-in-the-loop useless, requiring mitigation strategies and additional mechanisms to monitor the AI solution.
AUTOMATION COMPLACENCY	The deliberate preference for "false positives" or "false negatives" in the outputs generated by AI solutions can be harmful from an ethical standpoint, requiring proper justification not only during the development phase but also throughout the training and execution stages.
COMPENSATION FOR KNOWN BIAS	Human oversight can overcompensate for the errors and biases already identified in algorithm results, creating new distortions that affect the algorithm without a meaningful correlation to the databases.

SOURCE: MULLIGAN & BAMBERGER (2019); RUBENSTEIN (2021, AS CITED IN C4IR, 2022).

THE USE OF AI/ML SYSTEMS IN THE ADMINISTRATION OF HEALTHCARE FACILITIES

The documents on the use of AI/ML systems in the administration of healthcare facilities focus on two themes: (a) improving administrative activities and workflows, such as scheduling appointments, the optimization of hospital beds and operating rooms, and the evolution of patients throughout the different stages of the care process; and (b) using electronic devices, such as wearable devices, to create patient-centered care models.

In discussions on the first topic, one study worth highlighting is that of Sahni et al. (2023), who calculated the amount spent on health that could be saved annually in the United States by 2028 if existing AI technologies were disseminated among hospitals, doctors, and private clients. The conclusion is that the savings would be between 5% and 10% of current expenditures, or between US$ 200 billion and US$ 360 billion at 2019 values. Given this potential amount, the authors — affiliated to Harvard University and consulting company, McKinsey,

— conclude that AI technologies are not as widely used as they could be. They list the managerial challenges that could explain this phenomenon, divided into two groups. The first of these is made up of challenges faced by individual organizations in the healthcare sector, while the second comprises difficulties that, in order to be overcome, require the efforts of different stakeholders in the healthcare field. Frame 5 lists these challenges and concisely presents the considerations of Sahni et al. (2023).

FRAME 5 - CHALLENGES TO THE ADOPTION OF AI/ML SYSTEMS IN HEALTHCARE FACILITIES

MANAGEMENT (INTRA-ORGANIZATIONAL) CHALLENGES
Making it clear how the systems will add value to the organization's activities.
Attracting or developing professionals with the necessary skills to use the systems.
Assuring professionals at the healthcare frontline that the adoption of the systems will not harm patients.
Investing in factors that are critical to the functioning of the systems, such as databases.
Ensuring management of the data used from the outset of system use, in order to overcome problems, such as a lack of interoperability, fragmentation, and the preservation of privacy.
Creating governance and process models for the use of systems.
SECTORAL (INTER-ORGANIZATIONAL) CHALLENGES
Reducing the heterogeneity of EHR.
Increasing patients' confidence in the results generated by the systems.
Ensuring that systems are exposed to new data sets in order to keep them up to date.
Defining whether the time "saved" will be spent on new appointments or on non-clinical work activities, so that the systems lighten the workload of frontline healthcare professionals.
Ensuring that the systems have regulatory approval.

SOURCE: PREPARED BY THE AUTHOR BASED ON SAHNI ET AL. (2023).

Although the costs of AI/ML systems can also be an impediment to the adoption of the technology by healthcare facilities, there are few costs or savings estimates on the use

of AI in healthcare, and those that do exist focus on specific elements. According to Adler-Milstein et al. (2022), current information indicates that the global cost of developing and implementing AI/ML systems varies between US$ 15,000 and US$ 1 million. The authors also point out that:

> [...] another challenge is the tension between hiring a health care technology firm to develop or adapt the algorithms and tools into a health care environment versus hiring and supporting internal staff, which could cost between US$ 600 and US$ 1,550 a day. (Adler-Milstein et al., 2022, p. 49)

The use of electronic devices, such as wearable devices, to create patient-centered care models is discussed by Topol (2019) and Xie et al. (2021). Topol observes that the development of DL algorithms has focused on uses by healthcare professionals, with patient use taking a back seat.[14] Xie et al. (2021) discuss how AI and blockchain technologies can be integrated into wearable devices to manage chronic diseases. Although the authors argue that technological solutions of this kind can improve individual well-being, they also recognize that they are associated with different technical and social challenges, such as the security of accuracy rates, the interpretability of results, the interoperability of data needed for the technology to work properly, user protection, and the price of the technology.

Neither paper discusses the risk identified by Challen et al. (2019): Unscalable oversight, which is defined by the authors as the need for the user's constant attention to the AI/ML system. As an example, they cite autonomous subcutaneous insulin pumps that require the patient to provide exhaustive information about their diet so the pump can correctly adjust the level of the substance to be administered before each meal.

Hobensack et al. (2023), in turn, reviewed the literature on the use of ML techniques in electronic health data generated in home healthcare contexts for predicting adverse outcomes,

14 Junaid et al. (2022) mapped out different wearable devices that use ML techniques.

such as hospitalization. Both this work and that of Xie et al. (2021) mention the need to integrate the data produced by patient-centered devices into the functioning of health systems. They do not, however, explore the difficulties of doing so.

If left unaddressed, intra and inter-organizational challenges could result in AI/ML systems being underused in healthcare, thereby failing to contribute to reducing costs in the sector and, consequently, expanding access to quality services.

CONSOLIDATION OF THE RESULTS FOUND IN THE LITERATURE REVIEW

The papers discussed indicate that numerous socio-technical challenges need to be overcome so that AI/ML technologies can be useful in promoting predictive, preventive, personalized, and participatory health treatments. Ignoring them could lead to extreme situations, either a new "AI winter," in which the technology will be underused due to social fears, or a scenario in which AI/ML systems will be used in an undesirable way, i.e., without any respect for doctors and patients, or without knowing when and/or how they are useful for care processes.

In an effort to avoid either of these two outcomes, numerous researchers and institutions have proposed recommendations for the development and responsible use of AI in healthcare. The literature review identified numerous suggestions of this kind. This finding is echoed in the work of Jobin et al. (2019, as cited in Goirand, 2021): The researchers found that different actors linked to the development and use of AI technologies have jointly published at least 84 ethical frameworks in recent years. The expansion of publications of this type does not mean, however, that their recommendations are being internalized by healthcare actors. Evidence of this can be found in the work of Goirand et al. (2021), in which they analyzed 33 academic and non-academic documents published between 2015 and 2020 on the implementation of ethical frameworks in AI applications in healthcare. One of the main conclusions is that only eight documents mention any specific framework for ethics in AI.

The authors state that the result observed is due, among other factors, to difficulties in implementing an ethical framework for the specific situations in which AI systems are used. This finding indicates that tackling the obstacles and minimizing

the risks associated with such systems depend, albeit partially, on individual institutional efforts. It is true, however, that the success of these efforts depends on at least two factors: (a) the familiarization of the numerous actors linked to AI systems in healthcare with guidelines for the development and responsible use of this technology; and (b) the existence of public policies concerned with promoting this familiarization and facilitating institutional cooperation around the guidelines in question. Sections "Recommendations for the development and responsible use of AI: A review" and "Guidelines for the development and responsible use of AI in healthcare in Brazil: The role of the DHS" discuss how the WHO recommendations (2021) can be useful when formulating public policies aimed at these two objectives. This debate is especially relevant for Brazil since reflections on the relationship between AI and health are still in their infancy in the country: In the bibliographic universe on which this chapter is based, there are few works referring to Brazil.[15]

Although the term "ethics" recurs in the papers reviewed, it was decided to discuss the development and "responsible" use of AI in the Section "Recommendations for the development and responsible use of AI: A review." This semantic differentiation is due to the understanding that the second term gives equal prominence to the ethical and administrative challenges addressed, respectively, in the Subsections "The use of AI/ML systems in clinical practice," and "The use of AI/ML systems in the administration of healthcare facilities."

RECOMMENDATIONS FOR THE DEVELOPMENT AND RESPONSIBLE USE OF AI: A REVIEW

The WHO (2021) recommendations for the development and responsible use of AI are in line with the general challenge identified in the previous section, which is to develop public policies capable of facilitating and stimulating a positive relationship between AI and healthcare. For this reason, this section presents a general overview of the recommendations made by the institution, complementing them

15 Among the studies on Brazil, Dourado & Aith (2022) and Nunes et al. (2022) stand out.

with recommendations from other institutions and authors identified in the literature review. In the Section "Guidelines for the development and responsible use of AI in healthcare in Brazil: The role of the DHS," a comparison is drawn between these recommendations and the content of the Brazilian digital health strategy, the aim being to identify guidelines for the development of relevant public policies.

WHO RECOMMENDATIONS FOR THE DEVELOPMENT AND RESPONSIBLE USE OF AI

Published in 2021, the document, *Ethics and Governance of Artificial Intelligence for Health - WHO Guidance* (WHO, 2021), presents two sets of recommendations for the development and responsible use of AI technologies in health. The first comprises elements for the creation of national and international AI governance frameworks capable of enabling this technology to function as an ally in the construction of universal health cover. The second set of recommendations is aimed at AI system developers, ministries of health, and healthcare institutions, and comprises practical guidelines for these three stakeholders to address the challenges and risks of AI in health.

In an effort to summarize them, it can be stated that the two sets of recommendations have three common elements. The first of these is the protection of privacy, expressed in the ideas of privacy by design and privacy by default. Operationally, they mean that every possible effort must be made to ensure the privacy and confidentiality of the information used in the development, validation and use of AI technologies.

The relevance of public and private guidelines and official regulations that guide audits and risk assessments is the second element that runs through the WHO's different considerations (2021). The institution highlights the importance of official legislation, such as the General Data Protection Regulation, and international standards and codes of good practice, such as ISO standards, the guidelines of the US National Institute of Standards and Technology, the IEEE 7000 series, and Health Level 7. The WHO (2021) notes that the four standards are useful for promoting compliance and minimizing challenges linked to interoperability, with the first three referring to the

protection of privacy, and the fourth to the transfer of clinical and administrative health data.

Finally, the institution's recommendations make it clear that, both in the development and use of AI/ML systems in health, coordination between different stakeholders committed to enabling bottom-up evaluations is fundamental for minimizing risks and maximizing benefits. Bottom-up evaluations are more influential than evaluations conducted exclusively by public authorities.

Next, the challenges and risks discussed in Subsections "The use of AI/ML systems in clinical practice" and "The use of AI/ML systems in the administration of healthcare facilities" are revisited, and specific suggestions from the WHO (2021) for addressing them are presented. The recommendations regarding explainability are divided into "Ensuring the quality and representativeness of the data used" and "Developing systems that are attentive to the user experience."

Preserving patients' privacy. Consent is one of the main mechanisms for promoting the protection of privacy and confidentiality. In the light of the Brazilian General Data Protection Law (LGPD, 2018), it can be understood as a free, informed, and unequivocal manifestation by which the individual agrees to the collection and processing of their personal data for a specific purpose. However, numerous obstacles have been a challenge to its enforcement, such as the high frequency with which personal data has been collected, especially in the health area, and the clarity of the terms of consent. Aware of these challenges, the WHO (2021) cites three strategies to ensure consent: (a) electronic informed consent; (b) dynamic consent; and (c) broad consent.

In specific situations, consent can be an obstacle to the realization of social and collective benefits. It is up to government entities in such cases, not only to clarify why consent is contrary to the public interest, but also to articulate mechanisms that ensure the safe sharing of health data among different stakeholders. WHO (2021) mentions, for example, the data hub of the Department of Veterans Affairs and the Precision Medicine Initiative (All of Us) – both in the United States.

The institution also deals with strategies for preserving privacy and confidentiality when consent proves incapable of

doing so, such as anonymizing health data and utilizing federated data systems. These systems enable various institutions to apply the same ML model to their databases and compare findings across different contexts. In this way, data can be preserved where it is, without hindering the development of technology.[16] This is why international organizations, such as the World Economic Forum, consider such systems to be promising (WHO, 2021).

Ensuring the quality and representativeness of the data used. The risks described in Frame 3 are associated with the programming of AI/ML systems and become more likely as the quality of the underlying data worsens. For this reason, it is imperative that risk assessments are carried out at every stage in the development of the technology, and regularly once the technology is in use. The WHO (2021) also states that "AI technologies should be tested prospectively in randomized trials and not against existing laboratory datasets" (WHO, 2021, p. 141). The organization also notes that regulatory agencies can help, not only to ensure that these evaluations take place, but also that their results are clearly communicated.

Regarding the representativeness of the data used, the WHO (2021) presents a series of very specific recommendations, especially for the developers of AI/ML systems, which generally revolve around a common axis: "Examine the effects of ethnicity, age, race, gender and other traits, and ensure that AI technologies with biases do not have negative impacts on individuals and groups according to these different characteristics" (WHO, 2021, p. 137).[17]

Finally, the challenges to achieve explainability span "Challenge 2 - Ensuring the quality and representativeness of the data used" and "Challenge 4 - Developing systems attentive to the user experience". According to the World Health Organization (2021, p. 141):

> [...] Liability rules used in clinical care and medicine should be modified to assess and assign liability, including product liability, the personal liability

16 Rahman et al. (2022) analyze federated learning.

17 The WHO (2021) considerations on biased results suggest that algorithmic biases are negative when there are deleterious social implications.

> of decision-makers, input liability and liability to data donors. The rules should include causal responsibility, objective liability regimes and liability for retrospective harm as well as mechanisms for assigning vicarious liability when appropriate.

Definitions such as these must be coordinated by health ministries.

Ensuring the (data) infrastructure necessary for the technology to function properly. The WHO (2021) suggests that ministries of health assess whether the existing health structures in their countries are sufficient for operating, maintaining, and supervising AI/ML systems. If they are not, the organization points out that building links with civil society and international organizations could be essential for improving them. Regarding data infrastructure in particular, it is worth highlighting the recommendations of C4IR (2022): The requirement for open licenses and the use of free software can avoid technological dependence on specific suppliers (lock-in).

Developing systems that are attentive to the user experience. The intelligibility of the outputs from AI/ML systems is one of the main challenges in clinical practice for two reasons. First, frontline healthcare professionals need to trust the results generated by the technology; to do so they need to be sure that they can interpret them correctly. Second, the professionals in question need to be able to explain to patients how the results generated by the technology have informed their clinical conduct; therefore, it is essential that the developers of AI/ML systems try to ensure technical intelligibility in order to enable the interpretability and objective presentation of the results generated by the technology. For them to succeed in this dual mission, the WHO (2021) recommends that these developers engage different stakeholders in the development of AI/ML systems and seek to understand the contexts in which they will be used, since health-oriented AI technologies are dependent on the context in which they are employed.

Ensuring training for frontline healthcare professionals. There are various mentions of this challenge in the literature supporting this chapter. They are not accompanied,

however, by specific indications of how healthcare professionals could be trained to learn to interact with AI/ML systems. Nor is there any indication of the specific skills they would need to acquire in order to be able to use technological resources to improve service for citizens.

Making AI/ML systems affordable. The studies reviewed pay little attention to the fact that the costs of AI systems can be detrimental both to equal opportunities, by accentuating socio-economic inequalities, and to the potential of AI/ML systems for reducing the collective costs of healthcare services.

Defining how technology can improve clinical and administrative work. Among the management challenges listed in Frame 5, one is distinctly administrative in nature, i.e., demonstrating how AI/ML systems can add value to the activities of the organizations that intend to use them. This mission can be carried out by organizational leaders who, for each clinical and administrative situation, must assess whether the use of AI/ML systems is necessary and appropriate. The WHO (2021) lists a series of actions for these leaders to achieve the goal under discussion, such as comparing the risks and benefits of AI/ML systems with the risks and benefits of existing systems. It is also essential for the leaders of organizations to ascertain whether, in the specific local scenario in which they operate, the public they serve is in favor of using AI/ML systems in the treatment of their health problems. The WHO (2021) refers to this situation as the "social license" for the use of AI.

GUIDELINES FOR THE DEVELOPMENT AND RESPONSIBLE USE OF AI IN HEALTHCARE IN BRAZIL: THE ROLE OF THE DHS

The DHS is one of the main pillars of Brazilian digital health. Although it is not exclusively dedicated to AI, it could be a good starting point to guide the development and responsible use of this technology in healthcare in Brazil, because it is a state strategy that has been discussed and negotiated by different stakeholders. Published in 2020, it updates the Digital Health Action, Monitoring and Evaluation Plan (Plano de Ação, Monitoramento e Avaliação de Saúde Digital [PAM&A] 2019-2023) for Brazil.

The strategy comprises seven priorities, 18 sub-priorities and 36 strategic actions. They correspond to the guidelines, policies, ordinances, acts and initiatives approved within the scope of the Unified Health System (Sistema Único de Saúde [SUS]), as shown in Frame 6.

The seven challenges discussed in the Section "Recommendations for the development and responsible use of AI: A review" are restated and the points of contact between the WHO (2021) recommendations for each of them and the content of the DHS are discussed. Before doing so, however, it is worth mentioning that there are transversal axes in the DHS, one of which is the introduction of the Collaboration Space, understood as:

> [...] a conceptual, virtual, distributed, logical and physical space that enables collaboration between all stakeholders in Digital Health, with clear definitions of expectations, roles and responsibilities. The proposed collaboration is not exclusively technological and seeks to include models, services, methods and knowledge that are made possible or made more efficient by the use of Digital Health. (Ministry of Health, 2020, p. 14)

Both this transversal axis and the first priority of the DHS, as shown in Frame 6, are strongly aligned with the WHO's (2021) indication that there must be articulation between different stakeholders when developing, validating, and using AI/ML systems.

The WHO (2021) recommendations point to two other common elements. The first is the importance of public and private guidelines and official regulations that guide audits and risk assessments of AI/ML systems. This aspect is not foreign to the DHS, since manifestations to this effect can be found both in the strategy's references to the LGPD (2018), as discussed below, and in specific items of the strategic actions. The second topic common to the international organization's recommendations, discussed below, refers to the idea that privacy must be preserved at all costs, which is reflected in the ideas of privacy by design and privacy by default.

FRAME 6 - THE SEVEN PRIORITIES OF THE BRAZILIAN NATIONAL DIGITAL HEALTH STRATEGY 2020-2028

Governance and leadership for DHS	Ensuring that DHS28 is developed under the leadership of the Ministry of Health, but that at the same time it is capable of incorporating the active contributions of the actors that participate in collaboration platforms
The computerization of the three levels of care	Promoting the implementation of computerization policies for health systems by speeding up the adoption of electronic medical records and hospital management systems for integrating health services and processes
Support for improving attention to healthcare	Ensuring the National Health Data Network (NHDN) supports best clinical practices by way of services such as telehealth, apps developed by the Ministry of Health, and other apps that are developed by collaboration platforms
The user as protagonist	Engaging patients and citizens to promote their adoption of healthy habits and the management of their own health, that of their families, and of their community, and helping build the information systems that will be used
Preparing and training of human resources	Training healthcare professionals in health informatics, and ensuring recognition of health informatics as an area of research and a profession
The interconnectivity environment	Allowing NHDN to enhance collaborative work in all health sectors so that technologies, concepts, standards, service models, policies, and regulations are put into practice
Innovation ecosystem	Ensuring that there is an innovation ecosystem that makes the most of the interconnectivity environment in healthcare, thereby establishing itself as a large open innovation laboratory, subject to the guidelines, standards, and policies established by Priority 1

SOURCE: DHS (2020).

Preserving patients' privacy. The DHS emphasizes the importance of preserving privacy, a concern that is clearly expressed in its first priority, in which there are specific actions to ensure that the DHS's initiatives are aligned with the LGPD. In this sense, it highlights, for instance, the need to strengthen informed consent and improve data-sharing models. In this sense, the WHO (2021) discussions on consent models and federated data systems could be useful references for DHS efforts in relation to this priority.

Ensuring the quality and representativeness of the data used. The DHS does not explicitly address the challenges and risks linked to the (lack of) representativeness of health data. Attention to this topic is seen, however, in the strategy's emphasis on the quality of health data, which is manifested in its numerous strategic actions aimed at strengthening the RNDS. As part of the Connect SUS program, its main objective is "to promote the exchange of information between the points of the Healthcare Attention Network (RAS), allowing the transition and continuity of care in the public and private sectors" (Ministry of Health, 2020, p. 20).

The RNDS is a national platform that integrates health data from all over the country to enable interoperability between health information systems from all sectors. Such integration is strategic to the accumulation of clinical data from different sources and, consequently, to the formation of a large amount of individualized information that can be used to understand the health/disease situation of the population and to plan efforts, thereby offering results that can better guide decision-making by managers that will benefit communities and individuals. The fulfillment of this mission depends, among other elements, on the connection of the different health stakeholders to the RNDS. Even if this challenge is overcome, and the RNDS continues to consolidate,[18] it will not be enough to solve two other challenges linked to the quality of AI/ML systems: (a) the lack of consensus on how to report or compare the accuracy of AI/ML systems; and (b) the lack of testing in real clinical environments.

18 It is important to note that, during the COVID-19 pandemic, the RNDS showed important results, in a clear sign that it is consolidating.

The first of these can be overcome, at least in part, by alignment between the members of the Collaboration Space. Being directly linked to the debate on accountability, the second challenge demands clear attributions of responsibility for any errors in technology that may occur. Of a regulatory nature, this action requires building understanding between the different stakeholders involved in AI/ML systems; without this, regulation can be ineffective in protecting patients and healthcare professionals, or rigid to the point of hindering innovation. Because of its multi-sectoral nature, the Collaboration Space can be strategic for avoiding either of these two undesirable outcomes.

Ensuring the (data) infrastructure necessary for the technology to function properly. The DHS has various actions aimed at consolidating the RNDS that are strategic for building and maintaining the data infrastructure needed for the AI/ML systems to function properly. Three of them stand out: (a) the promotion of interoperability with various external systems, such as primary care systems, laboratories, and pharmacies, among others; (b) the implementation of a data repository to store health information; and (c) the adoption of internationally recognized and available standards for health information. While the first two strategic actions can avoid situations of technological dependence (lock-in), the third is essential for tackling the lack of standardization that characterizes many of the EHRs. On the other hand, it is less clear what elements of DHS could be mobilized to overcome another infrastructure challenge linked to the operation of AI/ML systems: The need for hardware with significant processing power. Little explored by the documents selected for this chapter, this challenge is discussed by Adler-Milstein et al. (2022).

Developing systems that are attentive to the user experience. Since its inception in the 1960s and 1970s, the field of health informatics has always argued that different social players should be involved in the development of technological systems aimed at health, especially their potential users. Conscious of this, the DHS presents a series of strategic actions that ensure centrality to citizens and health professionals in care processes. When observing the arrangement of the

priorities, it is noted that attention to citizens is intentionally placed at the center of DHS's seven priorities (as "The user as protagonist" is the fourth priority), suggesting that the first three and last three priorities should be directed towards them. However, since the seven priorities were developed before the global expansion in the use of AI/ML systems in health, the actions listed in the strategy contribute in a limited way for healthcare professionals and citizens to participate both in the development and evaluation of these systems' usage. The WHO (2021) encourages this type of participation.

Ensuring training for frontline healthcare professionals. The reviewed studies mention the need for healthcare professionals to be prepared to use AI/ML systems, however, they do not provide clear paths to achieve this objective. DHS has strategic actions that, if adapted to the particularities of AI, could be useful in filling this gap in the case of Brazil. Specific actions linked to the fifth priority ("Preparing and training of human resources") stand out: (a) raising and describing competencies, experiences, knowledge and skills associated with each functional profile needed for health professionals and managers, and information technology (IT) professionals, to be active participants in DHS; and (b) promoting recognition of health informatics as a profession in the Brazilian Classification of Occupations (CBO), which includes defining professional profiles and detailing their attributions, duties and ethical limits.

Making AI/ML systems affordable. The DHS includes actions such as mapping sources of public funding and establishing mechanisms for private funding, which include specific tasks, such as preparing the documentation necessary for accessing private resources. As indicated, the reviewed documents provide few indications on how to address the high costs of AI/ML systems, so enhancing the DHS's considerations on financing could be strategic, not only for developing national governance guidelines for AI in health but also for international guidelines. The Brazilian experience can be especially useful for other LMIC. The enhancement of the DHS's considerations on financing can be based on the sub-priority "value-based health" – linked to the seventh priority ("Innovation ecosystem") – since its

central objective is to encourage the testing of concepts, models, methods, and data sets to help overcome the challenge of measuring value in health.

Defining how technology can improve clinical and administrative work. This challenge intersects with others that focus on the education and training of professionals and making them ready to use AI/ML systems, albeit with a stronger emphasis on healthcare managers. An additional challenge they must face is the need to assess whether the institutions they are responsible for have a "social license" to use AI systems. This type of investigation is not simple. After all, local communities may oppose the use of technology, even though it can bring individual and collective benefits. Managers therefore need to be prepared to deal with this type of situation. Among the DHS mechanisms, the strategic actions of the third axis ("Support for improving attention healthcare") can be instrumental in leading community discussions on the potential benefits and risks of AI/ML systems.

CONCLUSION

Discussions about AI are often marked by temporal ambiguities that mix what technology may potentially do with what it currently accomplishes. (Meadows et al., 2020). Rather than discursive confusion, this mixing of tenses reveals that there is still uncertainty about the current stage of AI development. Without any pretense of offering a complete and definitive diagnosis of this situation, this chapter's aim was to map the challenges and opportunities envisioned for the adoption of AI tools in the healthcare sector, based on an analysis of publications on AI and healthcare. Elements were gathered that allow us to state that the potential of AI is still being explored in the field studied, and the realization of this potential depends on overcoming at least seven challenges: (a) preserving patients' privacy; (b) ensuring the quality and representativeness of the data used; (c) ensuring the (data) infrastructure necessary for the technology to function properly; (d) developing systems that are attentive to the user experience; (e) ensuring training for frontline healthcare professionals; (f) making AI/ML systems affordable; and (g) defining how technology can improve clinical and administrative work.

Next, efforts were made to understand how the seven challenges can be overcome, as well as to identify possible tools, policies, and guidelines in Brazil to help the country tackle them. Given this objective, DHS was analyzed to identify whether and how this strategy can guide responsible uses of AI in the country's healthcare sector. Actions that enable and facilitate the adoption of AI/ML techniques were highlighted, as were the potential adaptations in DHS that may be considered by professionals involved in caring for the health of the population.

The first group includes the existence of the Collaboration Space and the RNDS, as well as the DHS's considerations on value-based health, human resources training, and healthcare. Regarding the potential adaptations, at least five themes were identified: (a) strengthening the protection of patient privacy through specific measures, such as the adoption of federated systems and consent models specific to massive data collection; (b) the lack of consensus on how to report or even compare the accuracy of AI/ML systems; (c) the development of regulations that are neither ineffective when it comes to protecting patients and healthcare professionals, nor rigid to the point of hindering innovation; (d) the need for hardware with significant processing power; and (e) the participation of healthcare professionals and citizens in both the development and evaluation of the use of AI/ML systems.

In order for the changes imposed on healthcare by AI/ML systems to have positive results, it is necessary to formulate public policies that promote the development and responsible use of this technology in the country. However, these elements are not sufficient to achieve this goal; to attain it, it is necessary to better understand the current stage of the relationship between AI and healthcare in Brazil, which requires investigations that focus on the different stakeholders operating at the intersection of these two themes. The next chapters of this publication explore aspects presented here.

REFERENCES

Adler-Milstein, J., Aggarwal, N., Ahmed, M., Castner, J., Evans, B. J., Gonzalez, A. A., James, A. C., Lin, S., Mandl, K. D., Matheny, M. E., Sendak, M. P., Shachar, C., & Williams, A. (2022). Meeting the moment: Addressing barriers and facilitating clinical adoption of Artificial Intelligence in medical diagnosis. *NAM Perspectives, 2022*, 35-80. https://www.gao.gov/products/gao-22-104629

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L. Albahri, O. S., Alamoodi, A. H., Bai, J., Salhi, A., Satamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable Artificial Intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion, 96*, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. (2020). Explainability for Artificial Intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(310), 1-9. https://doi.org/10.1186/s12911-020-01332-6

Aquino, Y. S. J., Rogers, W. A., Braunack-Mayer, A., Frazer, H., Win, K. T., Houssami, N., Degeling, C., Semsarian, C., & Carter, S. M. (2023). Utopia versus dystopia: Professional perspectives on the impact of healthcare Artificial Intelligence on clinical roles and skills. *International Journal of Medical Informatics, 169*, 1-10. https://doi.org/10.1016/j.ijmedinf.2022.104903

Arbix, G. (2020). Transparency at the heart of building an ethical AI. *New Studies, 39*(02), 395-413. https://doi.org/10.25091/s01013300202000020008

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research, 22*(6), 1-7. https://doi.org/10.2196/15154

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of Artificial Intelligence-based FDA-approved medical devices and algorithms: An online database. *npj Digital Medicine*, *3*(118), 1-8. https://doi.org/10.1038/s41746-020-00324-0

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019). Hello, world: Artificial Intelligence and its use in the public sector. *OECD Working Papers on Public Governance, 36*, 1-185. https://dx.doi.org/10.1787/726fd39d-en

Bishara, A., Maze, E. H, & Maze, M. (2022). Considerations for the implementation of machine learning into acute care settings. *British Medical Bulletin, 141,* 15-32. https://doi.org/10.1093/bmb/ldac001

Centre for the Fourth Industrial Revolution Brazil. (2022). *Artificial Intelligence Public Procurement Guide*. https://c4ir.org.br/wp-content/uploads/2022/11/1648128585465GUIA-DE-CONTRATACOES-PUBLICAS-DE-AI_C4IR_v4.pdf

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial Intelligence, bias and clinical safety. *BMJ Quality and Safety, 28,* 231-237. https://qualitysafety.bmj.com/content/28/3/231

Daugherty, P. R., & Wilson, J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Review Press.

Dourado, D. A., & Aith, F. M. A. (2022). The regulation of Artificial Intelligence in health in Brazil begins with the General Personal Data Protection Law. *Revista de Saúde Pública, 56*(80), 1-7. https://doi.org/10.11606/s1518-8787.2022056004461

Du, Y., McNestry, C., Wei, L., Antoniadi, A. M., McAuliffe, F. M., & Mooney, C. (2023). Machine learning-based clinical decision support systems for pregnancy care: A systematic review. *International Journal of Medical Informatics, 173*, 1-9. https://doi.org/10.1016/j.ijmedinf.2023.105040

El-Sappagh, S., Ali, F., El-Masri, S., Kim, K., Ali, A., & Kwak, K-S. (2019). Mobile health technologies for diabetes mellitus: Current state and future challenges. *IEEE Access, 7*, 21917-21947. https://ieeexplore.ieee.org/document/8534339

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. This law addresses the processing of personal data, including in digital media, by natural persons or legal entities, whether public or private, with the aim of protecting the fundamental rights of freedom and privacy, and the free development of the personality of the natural person. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Ed.), *Media technologies: Essays on communication, materiality, and society* (pp. 167-193). The MIT Press.

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society, 38*, 549-563. https://doi.org/10.1007/s00146-022-01455-6

Goirand, M., Austin, E., & Clay-Williams, R. (2021). Implementing ethics in healthcare AI-based applications: A scoping review. *Science and Engineering Ethics, 27*(61). https://doi.org/10.1007/s11948-021-00336-3

Hobensack, M., Song, J., Scharp, D., Bowles, K. H., & Topaz, M. (2023). Machine learning applied to electronic health record data in home healthcare: A scoping review. *International Journal of Medical Informatics, 170*. https://doi.org/10.1016/j.ijmedinf.2022.104978

Hogg, H. D. J., Al-Zubaidy, M., Technology Enhanced Macular Services Study Reference Group, Talks, J., Denniston, A. K., Kelly, C. J, Malawana, J., Papoutsi, C., Teare, M. D., Keane, P. A., Beyer, F. R., & Maniatopoulos, G. (2023). Stakeholder perspectives of clinical Artificial Intelligence implementation: Systematic review of qualitative evidence. *Journal of Medical Internet Research, 25*. https://doi.org/10.2196/39742

Junaid, S. B., Imam, A. A., Balogun, A. O., Silva, L. C., Surakat, Y. A., Kumar, G., Abdulkarim, M., Shuaibu, A. N., Garba, A., Sahalu, Y., Mohammed, A., Mohammed, T. Y., Abdulkadir, B. A., Abba, A. A., Kakumi, N. A. I., & Mahamad, S. (2022). Recent advances in emerging technologies for healthcare management systems: A survey. *Healthcare (Switzerland), 10*(10), 1-45. https://doi.org/10.3390/healthcare10101940

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of Artificial Intelligence in medicine. *Gastrointestinal Endoscopy, 92*(4), 807-812. https://doi.org/10.1016/j.gie.2020.06.040

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. https://doi.org/10.1145/3375627.3375835

Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. *The Lancet Digital Health*, *4*(5), e384-397. https://doi.org/10.1016/S2589-7500(22)00003-6

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable Artificial Intelligence for healthcare: A systematic review of the last decade (2011-2022). *Computer Methods and Programs in Biomedicine*, *226*. https://doi.org/10.1016/j.cmpb.2022.107161

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy Artificial Intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 1-11. https://doi.org/10.1016/j.jbi.2020.103655

Meadows, R., Hine, C., & Suddaby, E. (2020). Conversational agents and the making of mental health recovery. *Digital Health*, *6*, 1-11. https://pubmed.ncbi.nlm.nih.gov/33282335/

Miailhe, N., & Hodes, C. (2017). Making the AI revolution work for everyone. *The Future Society, AI Initiative*, 1-29. https://thefuturesociety.org/wp-content/uploads/2019/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf

Ministry of Health (2020). *Brazilian National Digital Health Strategy 2020-2028*. https://bvsms.saude.gov.br/bvs/publicacoes/strategy_health_digital_brazilian.pdf

Moazemi, S., Vahdati, S., Li, J., Kalkhoff, S., Castano, L. J. V., Dewitz, B., Bibo, R., Sabouniaghdam, P., Tootooni, M. S., Bundschuh, R. A., Lichtenberg, A., Aubin, H., & Schmid, F. (2023). Artificial Intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1109411

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncology*, *20*(5), e262-e273. https://pubmed.ncbi.nlm.nih.gov/31044724/

Nunes, H. C., Guimarães, R. M. C., & Dadalto, L. (2022). Desafios bioéticos do uso da Inteligência Artificial em hospitais. *Revista de Bioética*, *30*(1), 82-93. https://doi.org/10.1590/1983-80422022301509PT

Organisation for Economic Co-operation and Development. (2024). *Recommendation of the council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pap, I. A., & Oniga, S. (2022). A review of converging technologies in eHealth pertaining to Artificial Intelligence. *International Journal of Environmental Research and Public Health*, *19*(18), 1-15. https://doi.org/10.3390/ijerph191811413

Rahman, A., Hossain, S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., Khan, S. I., Tiwari, P., & Band, S. (2022). Federated learning-based AI approaches in smart healthcare: Concepts, taxonomies, challenges and open issues. *Cluster Computing*, *26*, 2271-2311. https://doi.org/10.1007/s10586-022-03658-4

Sahni, N., Stein, G., Zemmel, R., & Cutler, D. (2023). The potential impact of Artificial Intelligence on healthcare spending. *National Bureau of Economic Research*, working paper 30857. https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges/potential-impact-artificial-intelligence-health-care-spending#:~:text=Yet%20healthcare%20lags%20other%20industries,billion%20annually%20in%202019%20dollars.

Schwalbe, N., & Wahl, B. (2020). Artificial Intelligence and the future of global health. *The Lancet, 395*, 1579-1586. https://doi.org/10.1016/S0140-6736(20)30226-9

Stanford University Human-Centered Artificial Intelligence. (2023). *Artificial Intelligence index report 2023*. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeke, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine*, *3*(17), 1-10. https://doi.org/10.1038/s41746-020-0221-y

Topol, E. (2019). High-performance medicine: The convergence of human and Artificial Intelligence. *Nature Medicine*, *25*, 44-56. https://doi.org/10.1038/s41591-018-0300-7

Tran, B. X., Vu, G. T., Ha, G. H., Vuong, Q-H., Ho, M-T., Vuong, T-T., La, V-P., Ho, M-T., Nghiem, K-C. P., Nguyen, H. L. T., Latkin, C. A., Tam, W. W. S., Cheung, N-M., Nguyen, H-K. T., Ho, C. S. H., & Ho, R. C. M. (2019). Global evolution of research in Artificial Intelligence in health and medicine: A bibliometric study. *Journal of Clinical Medicine, 8*(3), 360-378. https://pubmed.ncbi.nlm.nih.gov/30875745/

van Melle, W. (1978). MYCIN: A knowledge-based consultation program for infectious disease diagnosis. *International Journal of Man-Machine Studies, 10*(3), 313-322. https://doi.org/10.1016/S0020-7373(78)80049-2

World Health Organization. (2021). *Ethics and governance of Artificial Intelligence for health – WHO guidance*. https://www.who.int/publications/i/item/9789240029200

Xie, Y., Lu, L., Gao, F., He, S-J., Zhao, H-J., Fang, Y., Yang, J-M., An, Y., Ye, Z-W., & Dong, Z. (2021). Integration of Artificial Intelligence, blockchain, and wearable technology for chronic disease management: A new paradigm in smart healthcare. *Current Medical Science, 41*(6), 1123-1133. https://doi.org/10.1007/s11596-021-2485-0

Xu, Q., Xie, W., Liao, B., Hu, C., Qin, L., Yang, Z., Xiong, H., Lyu, Y., Zhou, Y., & Luo, A. (2023). Interpretability of clinical decision support systems based on Artificial Intelligence from technological and medical perspective: A systematic review. *Journal of Healthcare Engineering, 2023*, 1-13. https://doi.org/10.1155/2023/9919269

Yang, L., Ene, I. C., Belaghi, R. A., Koff, D., Stein, N., & Santaguida, P. L. (2021). Stakeholders' perspectives on the future of Artificial Intelligence in radiology: A scoping review. *European Radiology, 32*(3), 1477-1495. https://pubmed.ncbi.nlm.nih.gov/34545445/

Yu, K-H., Beam, A. L., & Kohane, I. S. (2018). Artificial Intelligence in healthcare. *Nature Biomedical Engineering, 2*, 719-731. https://doi.org/10.1038/s41551-018-0305-z

APPENDIX 1 - GENERAL INFORMATION ON THE STUDIES SELECTED FROM THE SCOPUS DATABASE

Nº	THEME	YEAR	AUTHORS	TITLE	JOURNAL
1	Clinical practice	2019	Kelly, C. J. et al.	*Key challenges for delivering clinical impact with Artificial Intelligence*	BMC Medicine
2	Clinical practice	2019	Challen, R. et al.	*Artificial Intelligence, bias and clinical safety*	BMJ Quality and Safety
3	Clinical practice	2019	Sheikhalishahi, S. et al.	*Natural language processing of clinical notes on chronic diseases: Systematic review*	JMIR Medical Informatics
4	Clinical practice	2019	Noorbakhsh-Sabet, S. et al.	*Artificial Intelligence transforms the future of health care*	American Journal of Medicine
5	Clinical practice	2019	Reddy, S. et al.	*Artificial Intelligence-enabled healthcare delivery*	Journal of the Royal Society of Medicine
6	Clinical practice	2020	Sutton, R. T. et al.	*An overview of clinical decision support systems: Benefits, risks and strategies for success*	npj Digital Medicine
7	Clinical practice	2020	Goecks, J. et al.	*How machine learning will transform biomedicine*	Cell
8	Clinical practice	2020	Ahmed, Z. et al.	*Artificial Intelligence with multi-functional machine learning platform development for better healthcare and precision medicine*	Database
9	Clinical practice	2020	Asan, O. et al.	*Artificial Intelligence and human trust in healthcare: Focus on clinicians*	Journal of Medical Internet Research
10	Clinical practice	2022	Loh, H. W. et al.	*Application of explainable Artificial Intelligence for healthcare: A systematic review of the last decade (2011-2022)*	Computer Methods and Programs in Biomedicine

11	Clinical practice	2022	Liu, X. et al.	*The medical algorithmic audit*	The Lancet Digital Health
12	Clinical practice	2022	Busnatu, S. et al.	*Clinical applications of Artificial Intelligence - An updated overview*	Journal of Clinical Medicine
13	Clinical practice	2022	Alanazi, A.	*Using machine learning for healthcare challenges and opportunities*	Informatics in Medicine Unlocked
14	Clinical practice	2022	Yang, L. et al.	*Stakeholders' perspectives on the future of Artificial Intelligence in radiology: A scoping review*	European Radiology
15	Clinical practice; Ethics and regulation	2022	Pap, I. A., & Oniga, S.	*A review of converging technologies in eHealth pertaining to Artificial Intelligence*	International Journal of Environmental Research and Public Health
16	Clinical practice	2022	Bishara, A. et al.	*Considerations for the implementation of machine learning into acute care settings*	British Medical Bulletin
17	Clinical practice	2022	Prakash, S. et al.	*Ethical conundrums in the application of Artificial Intelligence (AI) in healthcare - A scoping review of reviews*	Journal of Personalized Medicine
18	Clinical practice	2023	Moazemi, S. et al.	*Artificial Intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review*	Frontiers in Medicine
19	Clinical practice	2023	Xu, Q. et al.	*Interpretability of clinical decision support systems based on Artificial Intelligence from technological and medical perspective: A systematic review*	Journal of Healthcare Engineering
20	Clinical practice	2023	Hobensack, M. et al.	*Machine learning applied to electronic health record data in home healthcare: A scoping review*	International Journal of Medical Informatics

21	Clinical practice	2023	Du, Y. et al.	*Machine learning-based clinical decision support systems for pregnancy care: A systematic review*	International Journal of Medical Informatics
22	Management	2022	Rahman, A. et al.	*Federated learning-based AI approaches in smart healthcare: Concepts, taxonomies, challenges and open issues*	Cluster Computing
23	Management	2022	Junaid, S. B. et al.	*Recent advances in emerging technologies for healthcare management systems: A survey*	Healthcare (Switzerland)
24	Management	2022	Albalawi, U., & Mustafa, M.	*Current Artificial Intelligence (AI) techniques, challenges, and approaches in controlling and fighting COVID-19: A review*	International Journal of Environmental Research and Public Health
25	Management	2023	Luschi, A. et al.	*Semantic ontologies for complex healthcare structures: A scoping review*	IEEE Access
26	Management	2023	Vargas, V. B. et al.	*Influential factors for hospital management maturity models in a post - Covid-19 scenario - Systematic literature review*	Journal of Information Systems Engineering and Management
27	Ethics and regulation	2019	Wiens, J. et al.	*Do no harm: A roadmap for responsible machine learning for health care*	Nature Medicine
28	Ethics and regulation	2019	Watson, D. S. et al.	*Clinical applications of machine learning algorithms: Beyond the black box*	BMJ (Online)
29	Ethics and regulation	2019	Tran, B. X. et al.	*Global evolution of research in Artificial Intelligence in health and medicine: A bibliometric study*	Journal of Clinical Medicine
30	Ethics and regulation	2019	Guan, J.	*Artificial Intelligence in healthcare and medicine: Promises ethical challenges and governance*	Chinese Medical Sciences Journal

31	Ethics and regulation	2019	Ngiam, K. Y., & Khor, I. W.	*Big Data and machine learning algorithms for health-care delivery*	The Lancet Oncology
32	Ethics and regulation	2020	Amann, J. et al.	*Explainability for Artificial Intelligence in healthcare: A multidisciplinary perspective*	BMC Medical Informatics and Decision Making
33	Ethics and regulation	2020	Vollmer, S. et al.	*Machine learning and Artificial Intelligence research for patient benefit: 20 critical questions on transparency replicability ethics and effectiveness*	The BMJ
34	Ethics and regulation	2020	Char, D. S. et al.	*Identifying ethical considerations for machine learning healthcare applications*	American Journal of Bioethics
35	Ethics and regulation	2021	Goirand, M. et al.	*Implementing ethics in healthcare AI-based applications: A scoping review*	Science and Engineering Ethics
36	Ethics and regulation	2021	Markus, A. F. et al.	*The role of explainability in creating trustworthy Artificial Intelligence for health care: A comprehensive survey of the terminology design choices and evaluation strategies*	Journal of Biomedical Informatics
37	Ethics and regulation	2021	Saheb, T. et al.	*Mapping research strands of ethics of Artificial Intelligence in healthcare: A bibliometric and content analysis*	Computers in Biology and Medicine
38	Ethics and regulation	2022	Quinn, T. P. et al.	*The three ghosts of medical AI: Can the black-box present deliver?*	Artificial Intelligence in Medicine
39	Ethics and regulation	2022	Yoon, C. H. et al.	*Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned?*	Journal of Medical Ethics

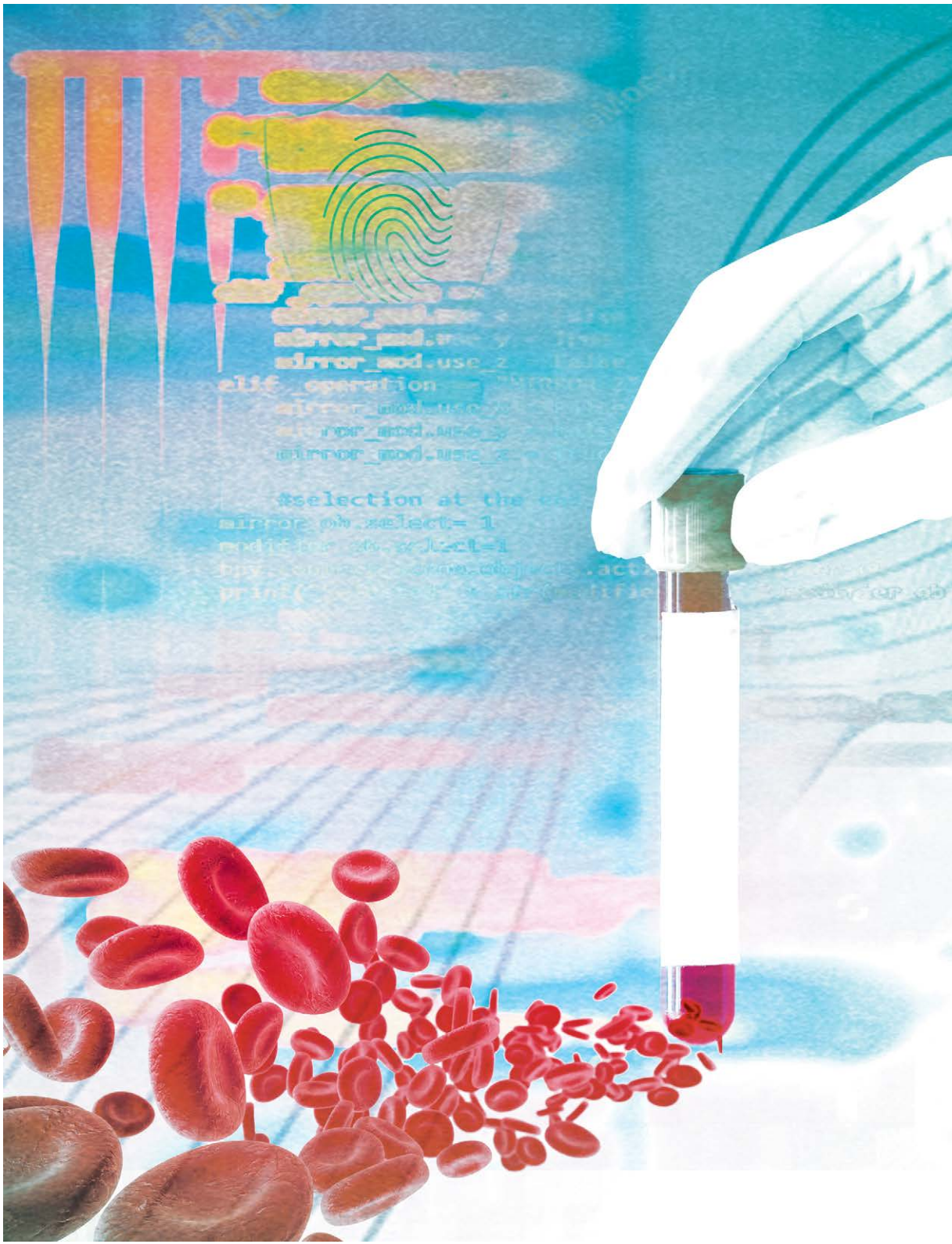
40	Ethics and regulation	2022	Salazar, L. H. A. et al.	*Application of machine learning techniques to predict a patient's no-show in the healthcare sector*	Future Internet
41	Ethics and regulation	2022	McLennan, S. et al.	*Embedded ethics: A proposal for integrating ethics into the development of medical AI*	BMC Medical Ethics
42	Ethics and regulation	2022	Angerschmid, Al. et al.	*Fairness and explanation in AI-informed decision making*	Machine Learning and Knowledge Extraction
43	Ethics and regulation	2022	Parviainen, J., & Rantala, J.	*Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care*	Medicine, Health Care and Philosophy
44	Ethics and regulation	2022	Bærøe, K. et al.	*Can medical algorithms be fair? Three ethical quandaries and one dilemma*	BMJ Health and Care Informatics
45	Ethics and regulation	2022	Siala, H., & Wang, Y.	*SHIFTing Artificial Intelligence to be responsible in healthcare: A systematic review*	Social Science and Medicine
46	Ethics and regulation	2022	Dourado, D. A., & Aith, F. M. A.	*The regulation of Artificial Intelligence for health in Brazil begins with the General Personal Data Protection Law*	Journal of public health
47	Ethics and regulation	2022	Nunes, H. C. et al.	*Bioethical challenges related to the use of Artificial Intelligence in hospitals*	Bioethics Magazine
48	Ethics and regulation	2022	artolovni, A. et al.	*Ethical, legal and social considerations of AI-based medical decision-support tools: A scoping review*	International Journal of Medical Informatics
49	Ethics and regulation	2022	Okolo, C. T.	*Optimizing human-centered AI for healthcare in the Global South*	Patterns

50	Ethics and regulation	2023	Giovanola, B., & Tiribelli, S.	*Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms*	AI and Society
51	Ethics and regulation	2023	Aquino, Y. S. J. et al.	*Utopia versus dystopia: Professional perspectives on the impact of healthcare Artificial Intelligence on clinical roles and skills*	International Journal of Medical Informatics
52	Ethics and regulation	2023	Albahri, A. S. et al.	*A systematic review of trustworthy and explainable Artificial Intelligence in healthcare: Assessment of quality bias risk and data fusion*	Information Fusion
53	Ethics and regulation	2023	Wu. C. et al.	*Public perceptions on the application of Artificial Intelligence in healthcare: A qualitative meta-synthesis*	BMJ open
54	Ethics and regulation	2023	Hogg, H. D. J. et al.	*Stakeholder perspectives of clinical Artificial Intelligence implementation: Systematic review of qualitative evidence*	Journal of Medical Internet Research
55	Ethics and regulation	2023	Sallam, M.	*ChatGPT utility in healthcare education research and practice: Systematic review on the promising perspectives and valid concerns*	*Healthcare (Switzerland)*
56	Ethics and regulation	2023	Harrer, S.	*Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine*	*eBioMedicine*
57	Ethics and regulation	2023	Lehoux, P. et al.	*Tools to foster responsibility in digital solutions that operate with or without Artificial Intelligence: A scoping review for health and innovation policymakers*	*International Journal of Medical Informatics*

Regulatory considerations on Artificial Intelligence for health[1]

World Health Organization and International Telecommunication Union

1 This chapter was adapted from the publication *Regulatory considerations on Artificial Intelligence for health* with the authorization of the World Health Organization (WHO) and the International Telecommunication Union (ITU). The adaptation and review of this text were not created by WHO. WHO is not responsible for the content or accuracy of this review. The original edition is the binding and authentic edition. The original publication is available at: https://iris.who.int/handle/10665/373421



The mission of the World Health Organization (WHO) is to promote health, keep the world safe, and serve the vulnerable, and it is articulated in its global strategy on digital health 2020–2025 (WHO, 2021a). At the heart of this strategy, WHO aims to improve health for everyone, everywhere by accelerating the development and adoption of appropriate, accessible, affordable, scalable, and sustainable person-centric digital health solutions in order to prevent, detect, and respond to epidemics and pandemics, developing infrastructure and applications. Many international organizations and global players are contributing to this area along with WHO, including the International Medical Device Regulators Forum (IMDRF), the Global Harmonization Working Party (GHWP), the United States Food and Drug Administration (US FDA), Health Canada, the International Coalition of Medicines Regulatory Authorities (ICMRA), the International Organization for Standardization (ISO), the Organisation for Economic Co-operation and Development (OECD), the United Kingdom of Great Britain and Northern Ireland's Medicines and Healthcare Products Regulatory Agency (MHRA), the South African Health Products Regulatory Authority (SAHPRA), the European Commission (EC), Singapore's Health Sciences Authority (HSA), the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), Japan's Pharmaceuticals and Medical Devices Agency (PMDA), Swissmedic and Australia's Therapeutic Goods Administration (TGA). These international and regional organizations and national authorities collectively recognize the potential of Artificial Intelligence (AI) in enhancing health outcomes by improving clinical trials, medical diagnosis and treatment, self-management of care, and personalized care, as well as by creating more evidence-based knowledge, skills, and competencies for professionals to support health care. Furthermore, with the increasing availability of healthcare data and the rapid progress of analytics techniques, AI has the potential to transform the health sector to meet a variety of stakeholders' needs in healthcare and therapeutic development.

In order to facilitate the safe and appropriate use of AI technologies for the development of AI systems in health care, the WHO and the International Telecommunication Union (ITU) have established a Focus Group on AI for Health (FG-AI4H). To support its work, FG-AI4H created several working groups, including a Working Group on Regulatory Considerations (WG-RC) on AI for Health. The WG-RC consists of members representing multiple stakeholders – including regulatory authorities, policymakers, academia, and industry – who explored regulatory and health technology assessment concepts and emerging "good practices" for the development and use of AI in health care and therapeutic development. The work of the WG-RC represents a multidisciplinary, international effort to increase dialogue and examine key concepts for the use of AI in health care.

This publication, which is based on the work of the WG-RC, aims to deliver an overview of regulatory considerations on AI for health that covers the following six general topic areas: Documentation and transparency, the total product lifecycle approach and risk management, intended use and analytical and clinical validation, data quality, privacy and data protection, and engagement and collaboration. This overview is not intended as guidance or a regulatory framework or policy. Rather, it is a discussion of key regulatory considerations and a resource that can be considered by all relevant stakeholders, including developers who are exploring and developing AI systems, regulators, and policymakers who are in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI-enabled medical devices, and health practitioners who deploy and use such medical devices and AI systems. Consequently, the WG-RC recommends that stakeholders take into account the following considerations as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development:

1. **Documentation and transparency**: Pre-specifying and documenting the intended medical purpose and development process, such as the selection and use of datasets, reference standards, parameters, metrics,

deviations from original plans, and updates during the phases of development, should be considered in a manner that allows for tracing the development steps as appropriate. A risk-based approach should also be considered for the level of documentation and record-keeping utilized for the development and validation of AI systems.

2. **Risk management and AI systems development lifecycle approaches**: A total product lifecycle approach should be considered throughout all phases in the life of an AI system, namely: Pre-market development management, post-market surveillance, and change management. In addition, it is essential to consider a risk management approach that addresses risks associated with AI systems, such as cybersecurity threats and vulnerabilities, underfitting, algorithmic bias, etc.
3. **Intended use, and analytical and clinical validation**: Initially, providing transparent documentation of the intended use of the AI system should be considered. Details of the training dataset composition underpinning an AI system, including size, setting and population, input and output data, and demographic composition, should be transparently documented and provided to users. In addition, it is key to consider demonstrating performance beyond the training and testing data through external analytical validation in an independent dataset. This external validation dataset should be representative of the population and setting in which it is intended to deploy the AI system and should be independent of the dataset used for developing the AI model during training and testing. Transparent documentation of the external dataset and performance metrics should be provided. Furthermore, it is important to consider a graded set of requirements for clinical validation based on risk. Randomized clinical trials are the gold standard for evaluating comparative clinical performance and could be appropriate for the highest-risk tools or where the highest standard of evidence is required. In other situations, prospective validation can be considered in a real-world deployment and implementation trial, which includes a relevant comparator that uses accepted groups. Finally,

a period of more intense post-deployment monitoring should be considered through post-market surveillance and market surveillance for AI systems.

4. **Data quality**: Developers should consider whether available data are of sufficient quality to support the development of the AI system to achieve the intended purpose. Furthermore, developers should consider deploying rigorous pre-release evaluations for AI systems to ensure that they will not amplify any of the issues discussed in Section "Data quality" of this document, such as biases and errors. Careful design or prompt troubleshooting can help identify data quality issues early and can prevent or mitigate possible resulting harm. Stakeholders should also consider mitigating data quality issues and the associated risks that arise in healthcare data, as well as continue to work to create data ecosystems to facilitate the sharing of good-quality data sources.
5. **Privacy and data protection**: Privacy and data protection should be considered during the design and deployment of AI systems. Early in the development process, developers should consider gaining a good understanding of applicable data protection regulations and privacy laws and should ensure that the development process meets or exceeds such legal requirements. It is also important to consider implementing a compliance program that addresses risks and ensures that the privacy and cybersecurity practices take into account potential harm, as well as the enforcement environment.
6. **Engagement and collaboration**: During the development of the AI innovation and deployment roadmap it is important to consider the development of accessible and informative platforms that facilitate engagement and collaboration among key stakeholders, where applicable and appropriate. It is fundamental to consider streamlining the oversight process for AI regulation through such engagement and collaboration in order to accelerate practice-changing advances in AI.

Finally, the WG-RC has provided a forum for regulators and subject matter experts to discuss regulatory considerations for

the use of AI technologies and development of AI systems for health and medical purposes. The WG-RC recognizes that the AI landscape is evolving rapidly and that the considerations in this deliverable may require expansion as technology and its uses develop. The working group recommends that stakeholders, including regulators and developers, continue to engage and that the community at large works towards shared understanding and mutual learning. In addition, established national and international groups, such as the International Medical Device Regulators Forum (IMDRF) and the International Coalition of Medicines Regulatory Authorities (ICMRA) should continue to work on topics of AI for potential regulatory convergence and harmonization.

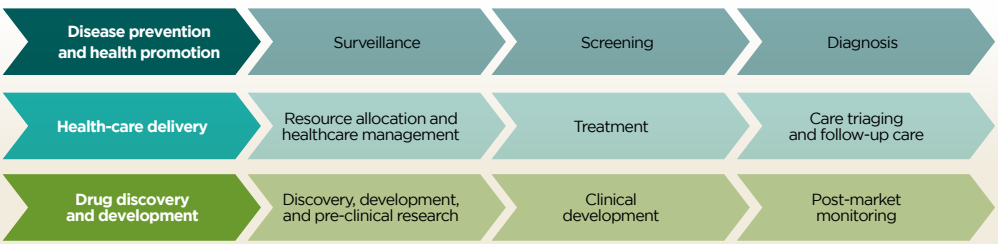
KEY ARTIFICIAL INTELLIGENCE APPLICATIONS IN HEALTH CARE AND THERAPEUTIC DEVELOPMENT

AI is increasingly being explored to advance health care on multiple fronts. The blending of technology and medicine in research and development is facilitating a wealth of innovation that continues to improve (Panesar, 2019). Many health-related AI systems already exist or are being developed to meet a variety of stakeholders' needs in health care and therapeutic development. These solutions have wide-ranging uses across the spectrum of healthcare delivery and therapeutic development. For instance, AI systems are being used in health care to support patients throughout the diagnosis and treatment of a disease, using solutions that support adherence to therapeutics and enhance communication capabilities with care providers.

Health care is becoming more patient-centric with personalized approaches to decision-making. This allows data to be used to improve patient and population wellness, patient education and engagement, prevention and prediction of diseases and care risks, medication adherence, disease management, disease reversal/remission, and individualization and personalization of treatment and care. Toward these ends, AI is increasingly being incorporated and utilized in the clinical setting. For instance, AI-enabled medical devices are being utilized to support clinical decision-making, and AI systems can facilitate clinical assessment of patients and care triaging. AI systems are also being used in the development and evaluation of medical products, including in drug

discovery to identify potential therapeutic candidates, and in clinical research for patient enrichment. Figure 1 illustrates areas of AI research and development across the spectrum of healthcare delivery and therapeutic development. The figure does not show an exhaustive listing of all AI applications but instead provides examples intended to illustrate the broad range of current and potential uses of AI systems.

FIGURE 1 - A GENERAL SPECTRUM OF AI RESEARCH AND DEVELOPMENT IN HEALTH-CARE DELIVERY AND THERAPEUTIC DEVELOPMENT



SOURCE: PREPARED BY THE AUTHOR.

The spectrum in Figure 1 assists in determining which regulatory considerations may be applicable and how they can be implemented. This document describes a selection of key regulatory considerations and discusses topic areas that are relevant to many stakeholders in the current AI for health ecosystem.

TOPIC AREAS OF REGULATORY CONSIDERATIONS

AI systems may be utilized across all aspects of health care and therapeutic development. Regardless of the category of the AI system application, regulators are keen to ensure not only that the AI systems are safe and effective for intended use but also that such promising tools reach those who need them as fast as possible. Dialogue among all stakeholders participating in the AI for health ecosystem, especially developers, manufacturers, regulators, users, and patients, is highly advised as the AI community matures. Consequently, this publication aims to establish a common understanding of the use of AI systems in health that can be relevant to stakeholders.

The subgroup leaders of the topic areas conducted a systematic literature review in 2020 of scientific publications in PubMed and other databases which included current guidelines and good practices in health care and therapeutic development. These sources informed the definition of the list of topic areas of regulatory considerations for the use of AI in health care and therapeutic development. At its first meeting, the WG-RC discussed the proposed topic areas and sought consensus to focus its deliverable on the six key areas listed in Table 1 while also discussing the remaining sections of this publication. The working group was divided into six subgroups composed of subject matter experts who drafted a section on each topic area.

The WG-RC stressed that this list is not a fully inclusive list of key considerations. The working group expects that the list will serve as a starting point for future deliberations and subsequent updates. For example, global systems for protecting intellectual property (IP) may be an important area to discuss as part of cross-jurisdiction regulations for some stakeholders (mainly AI system developers and manufacturers), and also in relation to, for instance, the protection of AI-related inventions by way of laws on patents and trade secrets. Although not addressed in this report, the World Intellectual Property Organization (WIPO) has already begun a dialogue on AI and IP (WIPO, n.d.). Thus, WHO will engage in this work together with WIPO and other relevant stakeholders.

TABLE 1 – SIX KEY TOPIC AREAS OF REGULATORY CONSIDERATIONS

TOPIC AREA 1	Documentation and transparency
TOPIC AREA 2	Risk management and AI systems development lifecycle approaches
TOPIC AREA 3	Intended use and analytical and clinical validation
TOPIC AREA 4	Data quality
TOPIC AREA 5	Privacy and data protect
TOPIC AREA 6	Engagement and collaboration

SOURCE: PREPARED BY THE AUTHORS.

DOCUMENTATION AND TRANSPARENCY

Documentation and transparency are critical concepts that are essential for facilitating scientific and regulatory assessments of AI systems. They also help ensure trust not only in the AI system itself, but also among developers, manufacturers, and end-users. Accurate and comprehensive documentation is essential to allowing a transparent evaluation of AI systems for health. This includes undertaking a total product lifecycle approach to pre-specifying and documenting processes, methods, resources, and decisions made in the initial conception, development, training, deployment, validation (data curation or model tuning), and post-deployment of health-related AI systems that may require regulatory oversight. The following discussion focuses on some elements related to documentation and transparency but is not fully inclusive of all the factors that are relevant to this important area.

Effective documentation and transparency help establish trust and guard against biases and data dredging. The same regulatory expectations and standards that ensure the safety and effectiveness of regulated products also apply to AI systems used in regulated areas. It is important for regulators to be able to trace back the development process and to have appropriate documentation of essential steps and decision points. For instance, aspects requiring careful documentation include specifying the problem that developers are attempting to address, the context in which the AI system is proposed to function, and the selection, curation, and processing of training datasets used in the development process.

Documentation should allow for the tracking, recording, and retention of records of essential steps and decisions, including justifications and reasoning for deviating from pre-specified plans. Effective documentation may also help to show that developers and manufacturers are taking into consideration the full complexity of the context within which the AI system is expected to operate. Moreover, developers and manufacturers should describe how the AI system is addressing the needs of users and why widening the user base would be appropriate. Without transparent documentation, it becomes hard to understand whether the proposed approaches will

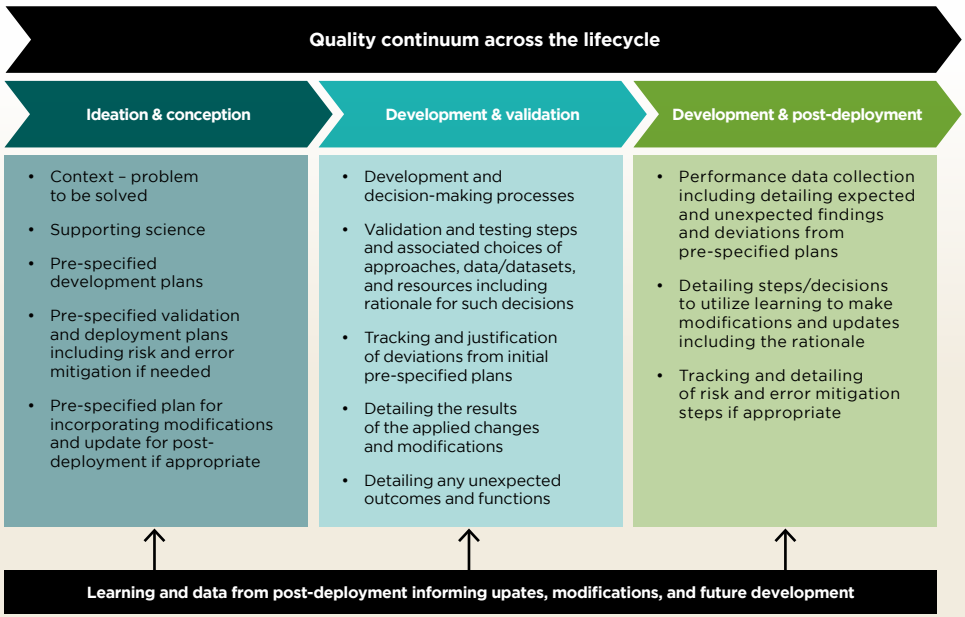
generalize from the retrospective clinical evidence presented in the regulatory submission to real-world deployments in new settings, which may markedly reduce performance (Wu et al., 2021). Figure 2 shows examples of essential steps and decision points that developers and manufacturers are encouraged to consider for documentation purposes.

Different entities with multidisciplinary expertise are likely to be involved in the development of AI systems for health and therapeutic development. There is a need to develop a shared understanding of procedures required for transparent documentation and to show that decisions are scientifically sound. Systems used to track and document the development processes and key decision points should record access and should be protected against data manipulation and adversarial attacks.

Documentation and transparency should not be seen as a burden but as an opportunity to show the strength of a science-based development that considers the full context in which the AI system is expected to be utilized, including the characteristics of end-users. Tools and processes for documentation should be proportional to the risks involved. Conversation with regulatory authorities prior to or in the early stages of development is encouraged and may provide vital help in informing documentation needs.

Beyond the regulatory perspective, it is important to note that effective documentation and other steps that help ensure transparency are important ways to establish trust and a shared understanding of AI systems in general. Steps to facilitate transparency include publishing in peer-reviewed journals; sharing data and datasets; and making code available to foster mutual learning and facilitate additional studies. Academic institutions, medical journals, regulatory organizations, and other stakeholders are working on advancing transparency for the use of AI in diagnostic and therapeutic development.

FIGURE 2 - EXAMPLES OF KEY DEVELOPMENT DECISION POINTS IN THE DEVELOPMENT OF AI SYSTEMS



SOURCE: PREPARED BY THE AUTHORS.

Collaborations, such as Consolidated Standards of Reporting Trials for AI (CONSORT-AI) (Liu et al., 2020) and Standard Protocol Items: Recommendations for Interventional Trials for AI (SPIRIT-AI) (Rivera et al., 2020), have given useful guidance about how to design studies to collect clinical evidence where AI systems are used, as well as how to publish the results. Transparency is not only an important consideration for building trust but can also be a useful tool for educating end-users. This can be achieved, if appropriate, by adapting communication strategies to the needs of end-users and other stakeholders such as patients and communities. As outlined in Figure 2, the development process of an AI system is multifaceted. A methodical approach to documentation throughout the full development cycle, including deployment and post-deployment, should be considered.

The following are some elements that might be useful to consider in terms of documentation and record retention.

Documentation across the total product lifecycle – ensuring a quality continuum

Developers should design, implement, and document approaches and methods to ensure a quality continuum across the development phases. Effective documentation outlining all phases of development would further enhance confidence in the AI system and would show how expected and unexpected challenges are identified and managed. Validation processes and benchmarking should be carefully documented, including the decisions for selecting specific datasets, reference standards, parameters, and metrics to justify such processes. For example, careful consideration should be given to documenting how and why specific data or datasets are selected to train, externally validate and retrain the model (e.g., post-deployment retraining).

Pre-specification and documenting the medical purpose, clinical context, and development

The intended medical purpose/function of the AI systems should be clearly documented. For instance, what is the problem that the AI system aims to resolve? This should take into consideration the full clinical and health contexts in which a tool is expected to function. For example, clinical care environments can be vastly complex and may involve several individuals with different roles and expectations. Documenting how the AI system should function in such active environments must be considered. As shown in Figure 3, there are multiple processes, testing/validation steps, and protocols that should be pre-specified and documented. Pre-specification is one of the most important elements that support trust and confidence in the development process. This will show evidence of a coherent development process and will be the basis for justifying any future changes.

Deployment and post-deployment

AI systems may be designed using data and datasets from specific populations. As with any therapeutics, once deployed, the AI systems will be utilized by a larger population and potentially variable end-users. Careful deployment plans

and justification for targeting different end-users should be considered. Manufacturers should be obliged to carry out post-market surveillance, which is the systematic process for collecting and analyzing experience gained from AI systems considered as medical devices that have been placed on the market (WHO, 2020). Deviations from pre-specified plans, updates, or changes to the AI system, post-deployment performance, data capture, and approaches to continued assessment of the system should also be documented. Such approaches will be increasingly relevant once there is a wider understanding that AI systems may change after deployment.

Risk-based approach and proportionality

Regulatory frameworks recommend a risk-based approach with processes in place to identify and mitigate errors, biases, and other risks in a manner proportional to their importance. A risk-proportional approach should also be considered for the level of documentation and record-keeping for AI systems. Developers of AI systems should keep in mind that regulatory organizations have avenues for dialogue and discussion that can be used to shed light on regulatory requirements.

RISK MANAGEMENT AND AI SYSTEMS DEVELOPMENT LIFECYCLE APPROACH

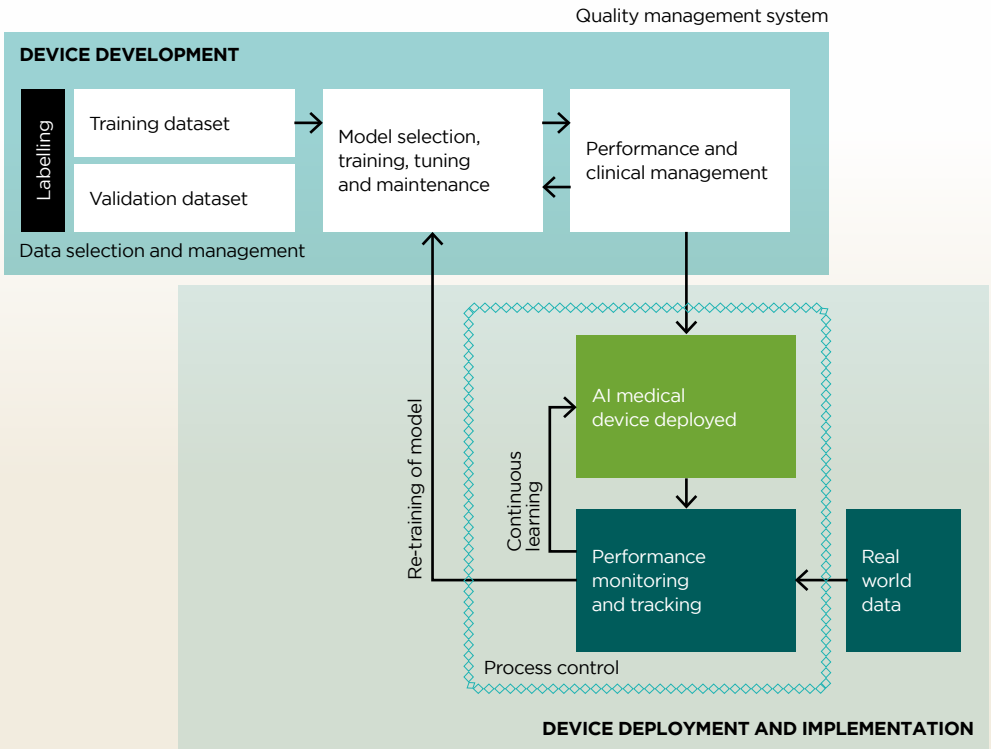
AI systems fall into many categories, e.g., devices that rely on AI and are used as medical devices (commonly known as Software as a Medical Device [SaMD]). Such categories of AI systems are defined by the IMDRF as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device" (IMDRF, 2013). However, the regulatory considerations for such a category of AI systems are similar to those of typical software that are regulated as medical devices, with the addition of considerations that may include continuous learning capabilities, the level of human intervention, training of models, and retraining (IMDRF, 2013). Furthermore, a holistic risk management approach that includes addressing risks associated with cybersecurity threats to an AI system, and the system's vulnerabilities, should be considered throughout the total product lifecycle. This topic area aims to present a

holistic risk-based approach to AI systems in general, and to those used as medical devices in particular, throughout their lifecycle, including during pre- and post-market deployment.

AI systems during the development and deployment process

Figure 3 illustrates the process of development and deployment of an AI system. Developers and implementers should establish measures to ensure responsible development of AI systems.

FIGURE 3 – THE PROCESS OF DEVELOPING AND DEPLOYMENT OF THE AI SYSTEM



SOURCE: HAS (2022).

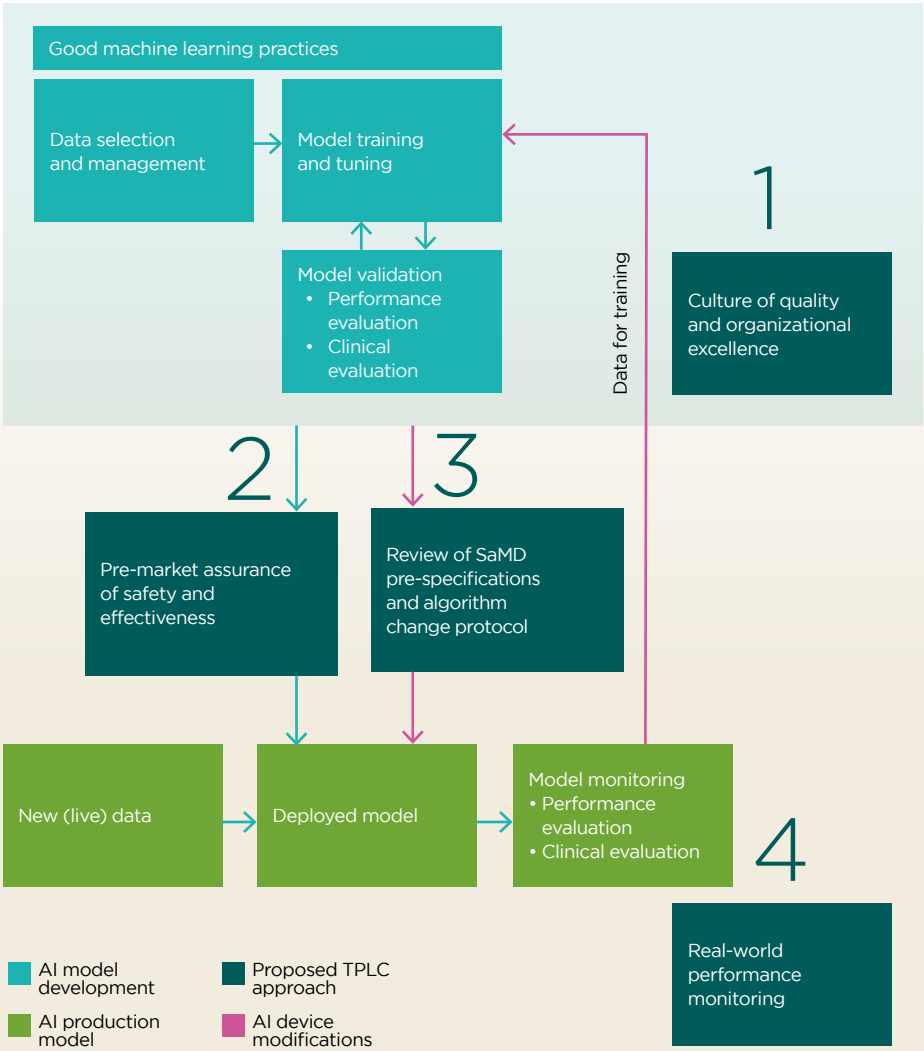
Figure 3 shows that all activities related to the design, development, training, validation, retraining, and deployment of AI systems should be performed and managed under a

quality management system based on ISO 13485 (HAS, 2022). For clinical endpoints, AI-specific monitoring dimensions include confidence (Oala et al., 2021), bias, and robustness (Oala et al., 2022).

AI systems development lifecycle

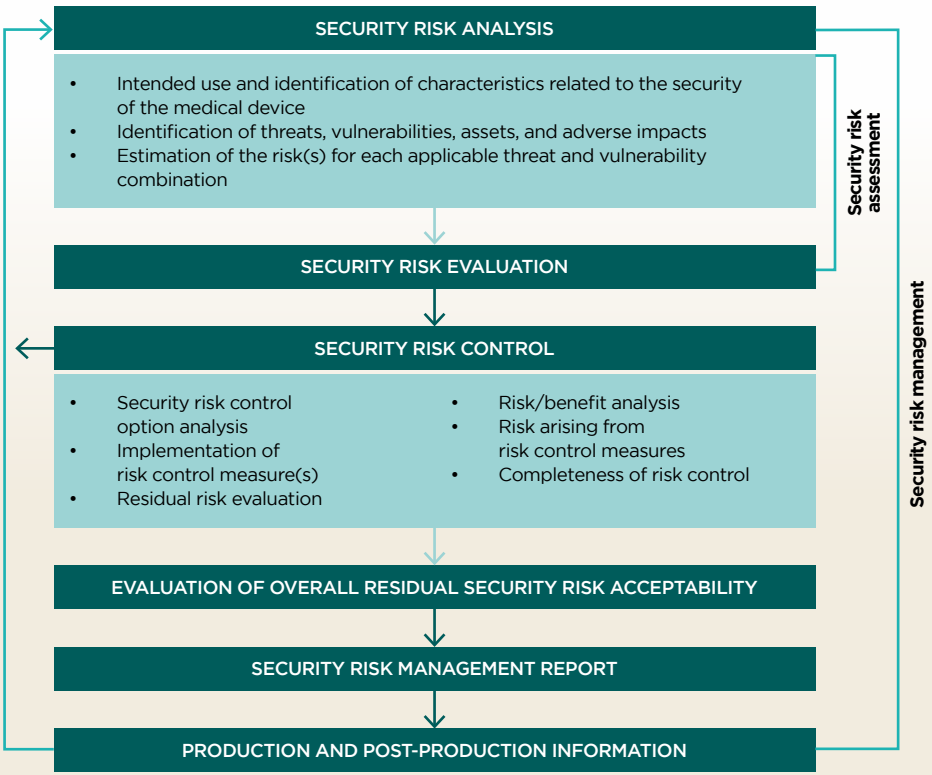
An AI system development lifecycle approach can facilitate continuous AI learning and product improvement while providing effective safeguards. This can be achieved if the development lifecycle approach involves appropriate development practices for the AI system. This approach could also potentially increase the trustworthiness, and assure performance and safety, of the AI system. An example is the Total Product Lifecycle (TPLC) approach (FDA, 2019) which could include the following four components (as illustrated in Figure 4):

FIGURE 4 - AI SYSTEM TOTAL PRODUCT LIFECYCLE APPROACH ON AI WORKFLOW



SOURCE: FDA (2019).

FIGURE 5 - IMDRF SCHEMATIC REPRESENTATION OF THE SECURITY RISK MANAGEMENT PROCESS



SOURCE: IMDRF (2019).

- demonstration of a culture of quality and organizational excellence of the manufacturer of the AI systems;
- pre-market assurance of safety and performance;
- review of AI systems' pre-specifications and algorithm change protocol; and
- real-world performance monitoring.

Holistic risk management

Holistic risk evaluation and management should be considered, taking into account the full context in which the AI system may be used. This could include not only the software or

AI system that is being developed but also other software that may be used within the same environment or context. Other risks, such as those associated with cybersecurity threats and vulnerabilities should be considered throughout all phases in the life of a medical device. Consequently, manufacturers of AI systems should employ a risk-based approach to ensure that the design and development of AI systems used as medical devices include appropriate cybersecurity protections. Doing so requires that manufacturers take a holistic approach to the cybersecurity of the device by assessing risks and mitigations throughout the AI system's development lifecycle. In order to achieve this, the IMDRF has published a security risk management process, as illustrated in Figure 5.

FIGURE 6 - GENERAL AI MEDICAL DEVICE RISK MANAGEMENT APPROACH

SOURCE: PREPARED BY THE AUTHORS.

However, to facilitate AI systems risk management, a general holistic management approach is introduced in this subsection with three broad management categories: Pre-market development management, post-market management, and change management. These categories are illustrated in Figure 6 and are discussed below:

- **Pre-market development management**: There is a need for transparency regarding the functioning of any manufactured AI-based devices to ensure that users can have a better understanding of the benefits, risks, and limitations of these AI-based systems (FDA, 2021). In addition, the controls and measures put in place to ensure that a developed AI system functions as expected while minimizing the risk of harm should be proportional to the risks that could occur if the AI system were to malfunction. For instance, failure of an AI system that is designed to encourage adherence to a healthy diet is

different from one that is designed to diagnose or treat certain diseases and pathologies. Therefore, developers should consider a risk-based approach through all processes to prioritize safety. Developers need to consider both the intended use of the AI system and the clinical context in order to evaluate the level of risk. For instance, the IMDRF risk framework for SaMD (IMDRF, 2014) identifies two major factors that may contribute to the impact or risk of an AI system. The first factor is the significance of the information provided by the AI system to the healthcare decision. The significance is determined by the intended use of the information, to treat or diagnose, to drive clinical management, or to inform clinical management. The second factor is the patient's healthcare situation or condition, which is determined by the intended user, disease or condition, and the intended population for the AI system, i.e. critical, serious or non-serious healthcare situations or conditions. Taken together, these factors related to the intended use can be used to place the AI system into one of four categories, ranging from lowest risk (I) to highest risk (IV) reflecting the risk associated with the clinical situation and device use.

The intended use and risk classification should be considered when testing different models and balancing trade-offs such as transparency and accuracy. In cases where training datasets are limited, simpler models, such as regression or decision-tree models, often provide equivalent or better results than more complex models and have the added benefit of more transparency and interpretability. On the other hand, in cases with larger and more complex datasets, complex models such as deep learning networks may not lend themselves to being explainable but may provide greater accuracy than simpler models. However, in cases in which there is a greater risk of harm, stakeholders should consider discussing the risks and benefits of choosing a more complex model and whether there are ways to mitigate the lack of interpretability and transparency and to build trust in the model through additional validation measures.

TABLE 2 - AI SYSTEMS RISK CLASSIFICATION

State of healthcare situation or condition	Significance of information provided by the AI system to the healthcare decision		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

SOURCE: IMDRF (2014).

Furthermore, depending on the level of risk, some AI systems may be approved as being available for full deployment whereas others may be initially authorized for deployment in more "AI-ready" institutions. "AI-ready" institutions are those that are certified on the basis of having stringent levels of surveillance in place with responsive backup systems to handle any failure of the algorithm in order to minimize the risk of patient harm.

Overall, it is important to achieve transparency among all AI-system stakeholders, including the developers, manufacturers, regulatory authorities, and implementers (i.e. users in healthcare settings, such as medical practitioners). Appropriate documentation of risk management and proper auditing procedures are examples of ways that help assure transparency. In general, auditing of specific key components of the AI medical device should be considered (e.g. certain software, hardware, training data, failure cases). For instance, it is important to do version control with training data because more data are added with each update. If an algorithm suddenly deteriorates in performance after an update, an inspection of everything that contributed to the update may be desired. In most cases, the element that will have changed is the addition of new training data by the developer (rather than changes to the software itself, such as modification to the neural networks). Moreover, given how unpredictable changes in performance can be for AI, active reporting and investigation of failure cases are recommended (as per the CONSORT-AI guidelines), although it is not prescriptive, given the wide range of available reporting and investigation avenues from

common-sense clinical auditing (i.e. human inspection) to technical solutions based on inference.

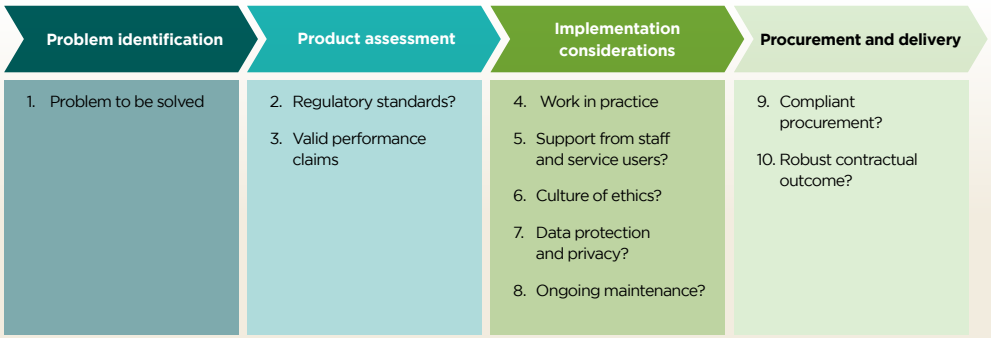
Although not specific to AI, there is a thickening web of country-, nation- and jurisdictional-specific legislations and laws that manufacturers and developers may need to consider for the development and deployment of regulated AI medical devices in health care. Such legislation includes the Personal Data Protection Act, Human Biomedical Research Act, Private Hospitals and Medical Clinics Act, Health Insurance Portability and Accountability Act and General Data Protection Regulation (GDPR). Compliance with relevant laws (local, cross- jurisdictional laws, and data protection acts) needs to be demonstrated by manufacturers and developers of medical devices whether they embed an AI component or not.

- **Post-market management**: Post-market monitoring and surveillance of AI medical devices allows timely identification of software- and hardware-related problems that may not be observed during the development, validation, and clinical evaluation of the device. New risks may surface when the software is implemented in a broader real-world context and is used by a diverse spectrum of users with different expertise. Companies involved in distributing AI medical devices (manufacturers, importers, wholesalers, authorized representatives, and registrants) are required to comply with their post-market duties and obligations which include reporting to relevant regulatory authorities in any of the following circumstances (WHO, 2020; HAS, 2022):
  - any serious public health threat;
  - death, serious deterioration in the state of health of the patient, user, or another person that has occurred;
  - death, serious deterioration in the state of health of the patient, user, or another person that may have occurred;
  - any field safety corrective action (such as return of a type of device to the manufacturer or its representative [also known as recall in some jurisdictions]; device modification; device exchange; device destruction; advice given by the manufacturer regarding the use of the device).

Furthermore, manufacturers should proactively collect information (through scientific literature and other information sources such as publicly accessible databases of regulatory authorities, user training and surveys) as part of their post-market surveillance plan. The plan should outline how manufacturers will actively monitor and respond to evolving and newly identified risks. Key considerations for the post-market surveillance plan include (HAS, 2022): Vulnerability disclosure, patching and updates, recovery, and information-sharing. Additionally, as part of the post-market duties and obligations, companies involved in distributing medical devices (manufacturers, importers, wholesalers, and registrants) are required to report adverse events associated with the use of software medical devices to relevant regulators.

In general, there is a need for both post-market clinical performance follow-up and periodical safety checks to report any potential harm. The intensity of post-market surveillance by the manufacturer may be risk- proportionate (according to consequences of failure [creating potential risk of harm] and likelihood of early detection of such failure). Finally, post-market surveillance requires a minimum level of evaluation for each site in order to ensure that potential algorithm vulnerabilities due to variation in local environments can be detected.

FIGURE 7 - THE UNITED KINGDOM'S NATIONAL HEALTH SERVICE - *A BUYER'S GUIDE TO AI IN HEALTH AND CARE*



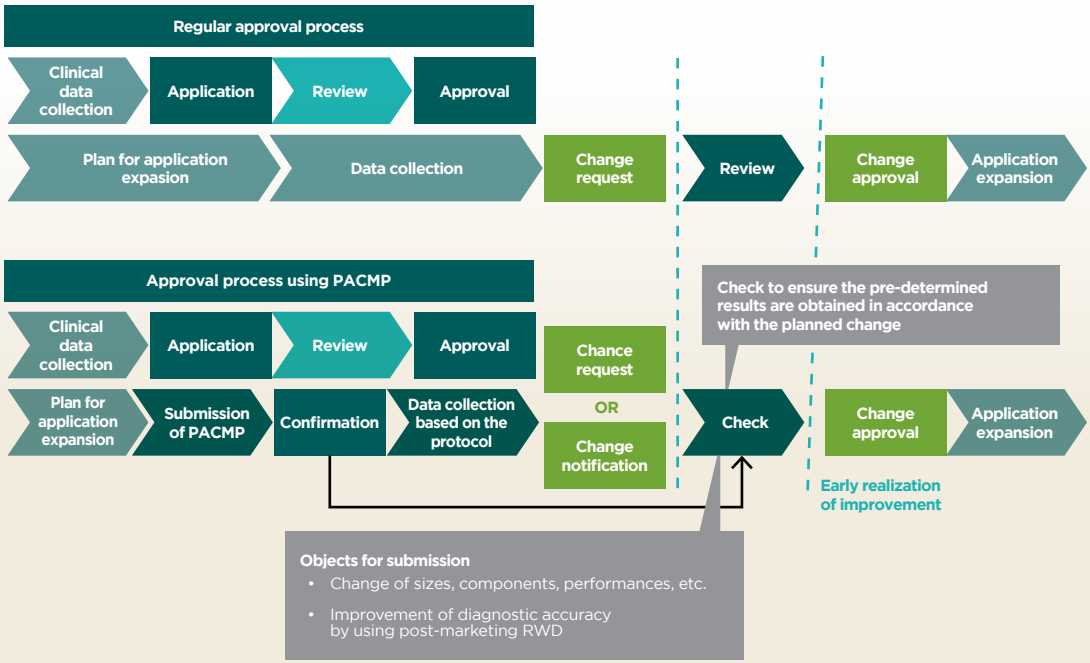
SOURCE: NHS (2020).

For example, the AI Lab of the National Health Service (NHS) in the United Kingdom of Great Britain and Northern Ireland published guidance to accelerate the safe and effective adoption of AI in health (NHS, 2022). The guide lists 10 questions in four categories to help buyers of AI products in order to make informed decisions, identify problems, assess products, and consider issues relating to implementation, procurement, and delivery (Figure 7).

- **Change management**: In view of the character of AI systems, it is important that the regulatory system enables continuous modifications for improvement to be made throughout the AI system's development lifecycle. The term "change" refers to such modifications, including those performed during maintenance.

  There are several proposed change management models and approaches for AI-based systems. Some consider change as part of the total development lifecycle (as in the TPLC approach) (FDA, 2019) (Figure 4). Other models focus on the change management process in the total lifecycle of medical device products which can be continuously improved. An example of this is the approach implemented by the Ministry of Health, Labor and Welfare of Japan and adapted in the Pharmaceuticals and Medical Devices Act as Post-Approval Change Management Protocol (PACMP) for medical devices (Pharmaceuticals and Medical Devices Agency Notification No. 14/2021) (Figure 8).

FIGURE 8 - POST-APPROVAL CHANGE MANAGEMENT PROTOCOL FOR MEDICAL DEVICES



SOURCE: PREPARED BY THE AUTHORS.

INTENDED USE AND ANALYTICAL AND CLINICAL VALIDATION

In principle, regulatory mechanisms are in place to answer the question: "Do the available data (included in the regulatory submission) support the conclusion that an investigational or experimental AI system is safe and performs sufficiently well to justify entry into the market and public access?" In addition to the principles discussed in "Documentation and transparency" and "Risk management and AI systems development lifecycle approach", one also must consider assessing if the use of the system is safe (i.e. it will not harm the user, the patient, or other persons) and if the claims made about its performance can be verified (see Figures 9 and 10). Evaluation of these claims for AI systems requires a clear use case description, demonstration of analytical and clinical validation, and assessment of the potential for bias or discrimination in the AI system.

Use case description, analytical and clinical validation

Clinical evaluation is the review of evidence that demonstrates the safety and performance of a given product for a given intended use. For AI systems (especially devices that rely on AI and are used for medical purposes), guidance is useful for collecting evidence of analytical and clinical validation. The performance of AI systems can be changed rapidly, not only as a result of a code change but also to provide different or additional training/tuning data. Consequently, clinical evaluation that accounts for total product lifecycle (TPLC) from development to analytical and clinical validation and to post-market surveillance should be considered for AI systems.

FIGURE 9 - DOMAINS OF HEALTH TECHNOLOGY REGULATION, ASSESSMENT AND MANAGEMENT FOR DRUGS AND DEVICES

SOURCE: PREPARED BY THE AUTHORS.

FIGURE 10 - IMDRF DESCRIPTION OF CLINICAL EVALUATION COMPONENTS

CLINICAL EVALUATION		
Valid clinical association	**Analytical validation**	**Clinical validation**
Is there a valid clinical association between your SaMD output and your SaMD's targeted clinical condition?	Does your SaMD correctly process input data to generate accurate, reliable, and precise output data?	Does the use of your SaMD's accurate, reliable, and precise output data achieve your intended purpose in your target population in the context of clinical care?

SOURCE: FDA (2019).

This topic area covers considerations related to use case descriptions (including statements of intended use) and analytical and clinical validation. These considerations follow the framework proposed by the WHO/ITU FG- AI4H Working Group on Clinical Evaluation (WG-CE) (ITU, 2020). A full description of this framework can be found in the deliverable for the WG-CE. The following section describes the key considerations and best practices and builds on the important work of national and regional regulatory authorities and international bodies such as the IMDRF. It is not intended to replace the work of these organizations. By outlining key considerations, this report draws attention to challenges that remain in this rapidly changing field. For instance, particular consideration is given to under-resourced settings that may have limited regulatory capacity at national level. The role of benchmarking in the evaluation of AI systems in health is also explored. Evaluation principles are applied to this topic area, and to the work of the WHO/ITU FG-AI4H in which benchmarking evaluation is a key component (ITU, 2021).

Intended use

AI systems are complex, dependent not only on the constituent code but also on the training data, clinical setting, and user interaction. They are often situated in a complex clinical pathway or are being introduced into new clinical pathways altogether (e.g. into new telemedical pathways or as part of new triage tools). Therefore, for AI systems, safety and performance can be highly context-dependent. The description of the use case has a substantial role both to inform end-users where the tool can be utilized safely and appropriately, and in regulated AI systems (the statement of intended use), to allow regulators to assess whether the evidence of the analytical and clinical validation steps is appropriate and sufficient for the intended use.

When developing a health-related AI system, it is important to describe the relevant use case. This consideration should cover the setting (geography, type of care facility), the population (ethnicity, race, gender, age, disease type, disease severity, co-morbidities) the intended user (healthcare provider or patient), and the clinical situation for which it is intended. Many

interventions, tests, and guidelines are prone to bias, and this is a particularly important consideration for AI systems which are highly sensitive to the characteristics of the data they were trained on and are prone to failure with unseen data types (such as a new disease feature or population type or context that was not previously encountered). Developers and manufacturers should also provide a clear clinical and scientific explanation of their tool's intended performance, including the populations and contexts for which it has been validated for use. Standardized reporting templates common to all stakeholders can help to communicate the intended use more effectively (Sendak et al., 2020; Verks & Oala, 2020; Oala et al., 2020). For some intended use cases there may be clear reasons why analytical performance of the tool would differ in different settings (Willis & Oala, 2021) (e.g. a symptom checker may perform differently in areas with a disease epidemiology that is different from the data on which it was trained). If this is the case, systematic known differences in performance should be included in the intended use statement. For other intended use cases, there may be emerging evidence that the tool under consideration, or another very similar tool, has been shown to have similar analytical performance in a wider setting than those in which the tool was initially developed and validated (Calderon-Ramirez & Oala, 2021) (e.g. retinal tools have been shown to have a similar performance in different populations (Bellemo et al., 2019)). Understanding of the generalizability of similar tools may be taken into account when providing a statement of the intended use or describing the use case (Mcdonald et al., 2021).

As part of the risk management process, regulators may wish to request evidence that developers have considered whether there are situations in which a tool should not be used (e.g. if there are insufficient training data for a particular patient group, or a lack of validation in a particular setting), or if there are potential risks associated with using the tools outside of the intended settings.

Analytical validation (also referred to as technical validation)

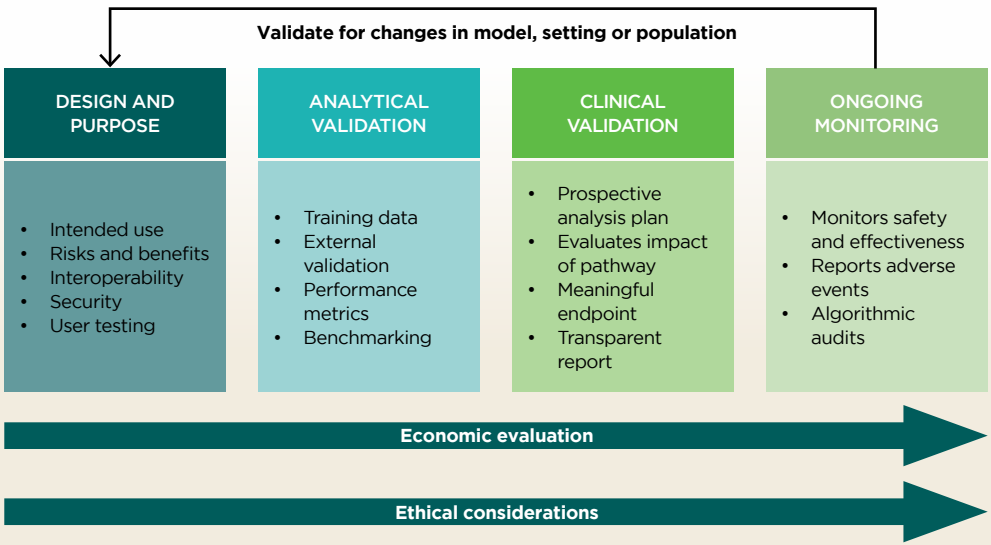
For the purposes of this document, analytical validation refers to the process of validating the AI system using data but without performing interventional or clinical studies. This may also be referred to as technical validation. Appropriate analytical validation demonstrates that a model is robust and performs to an acceptable level in the intended setting. It also enables the understanding of potential bias and generalizability (and any steps taken to understand these).

Developers and manufacturers should provide a description of the datasets used in the AI system's training, tuning, testing, and internal validation. The description of the intended use case (which can be on standardized reporting templates) should cover the size, setting, population demographics, intended user, and clinical situation (with input and output data). Transparency and documentation on dataset selection and characteristics are critical to ensure that AI systems are used appropriately. Developers and regulators may expect that the AI system has been externally validated in a dataset different from that in which it was trained and tested in order to demonstrate the model's external validity and generalizability beyond the original dataset. The external validation dataset is expected to be representative of the setting and population that are described in the intended use (gender, race, ethnicity) in order to demonstrate robust performance in the intended setting. The validation dataset should be of adequate quality, with appropriate robustness of labels. As part of the risk management process, it is important to identify any cases that are or may be high-risk (Oala et al., 2020).

Although bias, errors, and missing data are not unique to AI development, they are, nevertheless, serious concerns, which may arise for many reasons, including unequal and non-representative training or validation datasets, or structural bias in the systems where training data is generated (e.g. healthcare settings). Reporting the gender, race, and ethnicity of persons in the training and validation data cohorts, if feasible, can help to address the potential for bias and can avert its impact. For example, a better understanding of bias may help identify populations for which an AI system may

not function as expected. Post-market surveillance can also provide insights into the impact of potential bias.

FIGURE 11 - OVERVIEW OF FRAMEWORK FOR CLINICAL EVALUATION OF AI MODELS IN HEALTH DEVELOPED BY THE WG-CLINICAL EVALUATION



SOURCE: PREPARED BY THE AUTHORS.

Obtaining datasets for training, testing and validation that are sufficiently representative and of sufficient quality can be difficult. Local, regional and national bodies interested in procuring AI systems could hold their own hidden dataset to enable external validation (e.g. a recent scheme of the United Kingdom of Great Britain and Northern Ireland's NHSX has nationally representative datasets for some common use cases). Access to representative datasets for validation is a particular concern in many low- and middle-income countries. Where datasets are available in low-resource settings, there may also be limitations in the quality of the data. The ability to produce robust datasets with high-quality ground truth labels is likely to be affected by limitations elsewhere in the health setting where there may be barriers that impede access to diagnosis and treatment. These major challenges, which

have the potential not only to propagate inequality of access but also to compromise safety and performance of AI-based tools, are potential areas for future work. In this regard, the newly launched International Digital Health & AI Research Collaborative (iDAIR)[2] notes that collaborative, distributed, and responsible use of data is at the heart of its strategic plan.

While most regulatory agencies have national or regional remits, some countries with limited regulatory capacity tend to rely on decisions made by other major regulators. The availability of independent, hidden, representative datasets also offer particular advantages to countries that do not have their own regulatory process, or where regulatory decisions may be informed by dossiers provided to other bodies. However, the performance of AI-based systems is highly dependent on the context. In order to rely on regulatory review and decisions, many regulators (whether national or regional) could perform analytical validation as a second local validation step to ensure that the performance metrics obtained are consistent with those demonstrated in other regulatory jurisdictions. This could be best prioritized through a needs-based approach, e.g. the identification of key areas in which AI-based tools are promising and could provide local value, and the potential prospective creation of datasets to support validation.

In order to understand the performance of an AI system, an evaluation against an accepted standard should be made. The most appropriate standard for comparison may differ by intended use but commonly used standards are human performance in a similar task or other models (e.g. derived from logistic regression) with strong evidence-based or mandated standards of accuracy, sensitivity, and specificity (such as for screening tools). Depending on the intended use case, the requirement for comparative performance may be more or less stringent (e.g. when used as a triage or screening tool, a different level of comparative performance may be acceptable compared to a tool used for diagnosis).

2 Find out more: http://i-dair.org

Some limited comparative benchmarking of AI systems has been performed in a single setting but may become more common as the number of available tools increases (Salim et al., 2020). In the future, if an AI system has proven clinical efficacy and safety in a particular setting, it may be possible and appropriate to benchmark other newer tools against that AI system to understand potential similarities in performance. Benchmarking software is being developed as part of the work of the Open Code Initiative (ITU, 2022). Platforms such as this may also be useful to perform repeated algorithmic validation of models that have been updated. However, this is currently not the case for any use cases, and benchmarking thus far has been used only to understand comparative analytical performance. In addition, repeatedly using the same data for benchmarking multiple updated models (and thus, even if inadvertently, for training the test) risks introducing bias, and this should be taken into account when benchmarking is considered.

A designated FG-AI4H working group on data and AI solution assessment methods[3] provides guidance on the methods, processes and software development for the analytical validation of health-related AI systems (Oala et al. 2020).

Clinical validation

Analytical validation performed retrospectively on an existing dataset provides measures of performance (accuracy, negative predictive value, positive predictive value) but does not allow for evaluation of other factors that may affect the tool's performance (e.g. user interaction, workflow integration, and unintended consequences of the tool within a complex clinical pathway).

Both national and international bodies have proposed a graded set of requirements based on risk for digital health tools (including significance of the information provided by the tool and the state of the health condition) (IMDRF, 2016; National Institute for Health and Care Excellence [NICE], 2019). The IMDRF document on clinical evaluation of SaMD

3 Find out more: aiaudit.org

(Table 2) (IMDRF, 2014) proposes that devices in category I are the lowest-risk tools that have evidence of analytical validity, and that a novel tool in this category would require manufacturers to collect real-world performance data and generate a demonstration of scientific validity. For higher-risk SaMD, clinical evaluation evidence is expected on the basis of evidence of analytical validity. There is no universal agreement on the appropriate level of evidence of adequate clinical performance for a novel AI tool before deployment and this is the subject of a separate working group within the FG-AI4H (WG-CE).

Randomized clinical trial data are the gold standard evaluation of comparative clinical performance and may be appropriate for the highest-risk devices where an AI tool has no demonstrated performance in that setting, or for large national procurement bodies that seek evaluation of performance before national expenditure. A trial that is expected to guide clinical practice should have a clinically meaningful primary endpoint (morbidity, mortality) but, in certain situations, the event rate or time lag between the trial and the endpoint may result in a more feasible surrogate endpoint. Reporting guidelines backed by the widely accepted EQUATOR network are now available for protocols and clinical trials using AI systems (Liu et al., 2020). However, there are currently a small number of actively recruiting or completed randomized trials in this field (Topol, 2020).

Randomized clinical trials have potential limitations that may make other options preferable (trials can be slow, or expensive, and may evaluate performance in specific groups under trial conditions). Where randomized evidence may not be necessary (e.g. the evidence required may be proportional to the risk or cost of a tool), prospective validation in a real-world deployment and implementation trial, with a relevant comparison group showing improvement in meaningful outcomes using validated tools or widely accepted and verified endpoints and with systematic safety reporting, may be appropriate. Clinical performance should be considered in the context of the capabilities of health workers, available Internet bandwidth and health informatics infrastructure, and real-time data pipelines. Developers should provide a

description of the steps taken to perform clinical validation in a context similar to that available in the intended use setting.

Further consideration of the most appropriate level or type of clinical evaluation for a digital health intervention will be provided by the WG-CE.

In some situations, as described below, special considerations apply. For instance:

Post-market monitoring

Post-market monitoring in some regulatory contexts relies heavily on reporting of adverse events. Recent WHO guidance recommends that proactive post-market surveillance must be carried out by the manufacturer.

As part of a TPLC approach to regulation in this context, further prospective post-market clinical follow-up should be completed after deployment. Regulators must be notified of reportable incidents (adverse events), and findings from more continuous monitoring using real-world data may help developers and regulators better understand and assure the safety and performance of these devices in real-world use. For prospective monitoring of real-world data, significant investment will be required in prospectively curating and labeling validation data. A defined period of close monitoring may be appropriate for AI-based tools for those with high risk given their tendency to overfit on erroneous data features and produce unpredictable errors on unseen data features combined with the lack of data from use in real-world settings with long-term results. Regulators may recommend that manufacturers develop specific market surveillance measures that are appropriate for AI systems.

Changes to the AI tool

An update of an AI tool by a change of code, change of the user interface, or provision of further training data may alter the analytical or clinical performance of an AI system. The group is not aware of currently approved medical AI systems that are "continuously learning" but anticipates that these may be developed. Such AI systems would require a risk-benefit evaluation in keeping with the concepts in this document and with the clinical evaluation of AI systems for health. Taking

"checkpoints," by evaluating the tool as it is currently performing at regular intervals, enables regular evaluation and could signal changes in performance. Depending on the risk of the AI systems and the extent of the changes, appropriate validation must be agreed by the developer and the regulator. Analytical validation against previously unseen datasets, or benchmarking against approved datasets representative of the intended setting or population, could be useful in this scenario.

Low- and middle-income countries

There is considerable variation in the regulatory implementation of medical devices, and therefore also in deployed AI technologies and developed AI systems. Some countries lack a dedicated national regulatory body. The WG-RC meetings have provided a forum for the sharing of expertise and discussion of common problems, including for regulatory bodies and other interested stakeholders, some of whom have aligned remits. Furthermore, there are important regulatory considerations related to the intended use and analytical and clinical validation of AI systems in health. First, in low- and middle-income countries, one of the potential uses of AI technologies is in bringing specialized AI-based systems or knowledge to areas which do not have a relevant medical specialist (e.g. interpreting retinal scans, histopathology slides, or radiology images). In high-income countries, AI systems are more often positioned as an adjunct to medical professionals. Using an evaluation performed to support regulation in a high-income setting to inform how such AI systems are used in low- or middle-income settings may, therefore, not be appropriate. Thus, the full context of healthcare infrastructure and resources should be considered. Second, some regulatory bodies rely on decisions from other bodies to support their regulatory work. Given that the performance of AI systems may be highly context-dependent, additional steps may be required. There is a concern that developers may not ensure adaptation or evaluation for resource-limited settings if the market there is less attractive. Regulatory agencies in high-income countries could support this adaptation, which could also increase the generalizability and robustness of AI systems. However, this would require adaptive studies

to ensure wider use in low- and middle-income countries or the use of incentives to encourage additional development, testing, and validation. The availability of a range of representative datasets would support local analytical validation. Finally, AI systems for health can be highly sensitive to shifts in data distribution and features. They may, therefore, be sensitive to differences in disease prevalence when moving from high-income to low-income countries, with the possibility of lower performance without appropriate evaluation or tuning with local data.

DATA QUALITY

Data in current health ecosystems

The health sector has been very receptive to the benefits of AI thanks to the explosion of data and accessibility to computational power. Data are the most important ingredient for training AI/ML algorithms and can be classified on the basis of format, structure, volume, and many other factors. Data can take any form, including character, text, words, numbers, pictures, sound, or video. Also, these data can be structured, semi-structured, or unstructured (Panesar, 2019). Structured data are normally stored in databases that are structured in a manner that follows a specific model or scheme, such as data stored in electronic medical records, mobile devices, and Internet of Things (IoT) devices. Regardless of the format, structure, or volume of the data, a more general classification can be based on the following 10 Vs of data (Panesar, 2019) (as illustrated in Figure 12): Volume, veracity, validity, vocabulary, velocity, vagueness, variability, venue, variety, and value.

Good quality data in health AI systems

All AI tasks and solutions use some form of data, regardless of their characteristics, to facilitate machines to learn, adapt, and improve their learning. However, data quality greatly influences the success of such solutions' safety and effectiveness. "Good-quality data" is an ambiguous term that is open to misinterpretation. Therefore, gaining a good understanding of the datasets used, for example, from the 10 Vs perspective is crucial to assess data quality in AI systems during development and even afterwards. Section "Key quality data challenges and

considerations for health AI systems" highlights key challenges and considerations for all stakeholders, including developers and regulators, when handling data in AI systems in order to achieve good data quality.

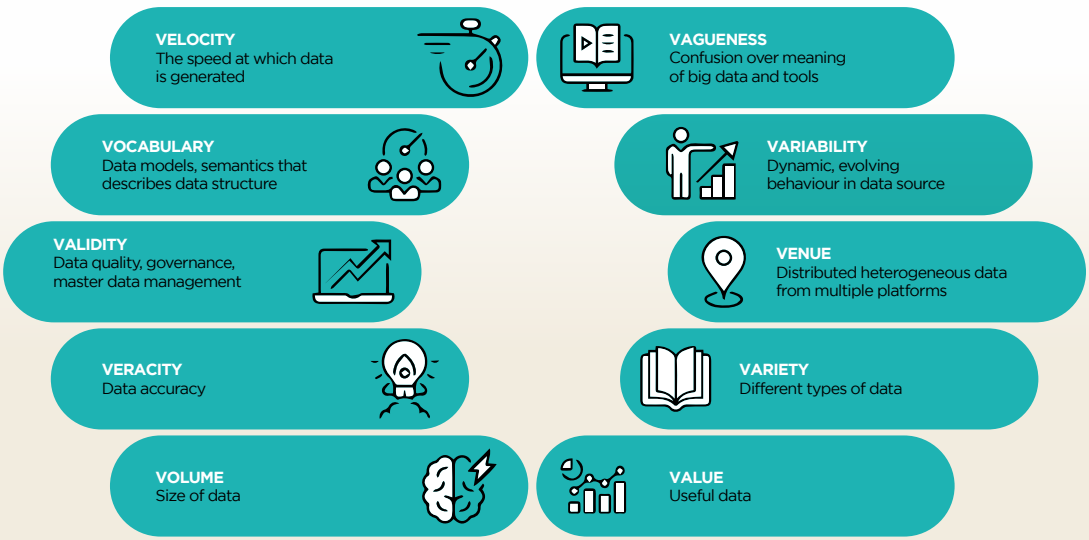
Key quality data challenges and considerations for health AI systems

The availability of good-quality datasets that are clinically relevant is one of the key challenges that developers face. However, data of varying quality can still be used depending on the purpose, and thus developers should determine if available data are of sufficient quality to support the development of systems that can achieve their intended goal. The lack of good-quality datasets for use in the development of AI systems may hinder their effectiveness and potential benefits. Data that are not of sufficient quality for the intended purpose can also lead to many problems, such as bias and errors. Some data quality issues that often arise when developing AI systems, and that need to be considered by all stakeholders, are discussed in this section and summarized in Table 3. These issues and considerations can relate directly to dataset management, the machine learning (ML) model, the infrastructure used to manage the data, or general governance aspects, as follows:

- **Dataset management**: When managing datasets for ML models, a clear data management plan should be pre-specified and well documented. Data management approaches should be risk-based and fit for purpose. This may include data selection volume (including volume of data used and volume of available data), splitting, cleansing (including any AI algorithms that were used to clean the data), data usability (including how well the dataset is structured in a machine-readable format), labelling, dependencies, augmentation and streaming. If data augmentation is relevant, it is important to develop a clear data augmentation strategy. The developers should also consider putting in place good data accountability practices for those handling the data in order to ensure that data quality and integrity are maintained throughout the lineage of the data. This is also essential

for knowledge management and transfer in a highly evolving field. Further, in addition to the handling of the data, the capacity to plan for and conduct data analyses is also important.

FIGURE 12 - THE 10 VS OF DATA



SOURCE: PANESAR (2019).

- **Data inconsistency**: High heterogeneity in the syntax of the data may require harmonization in order to address issues related to multiple data sources with varying standards, formats, schemas, structures, and ambiguous semantics and generate a coherent dataset for the purpose of comprehensive analysis, which is especially challenging when using healthcare data. For instance, much of the data collected from various information silos is inconsistent, incompatible, or not executable in machine-readable formats. For multiple data sources, there may be variations in how the data are captured (e.g. definitions of individual variables).
- **Dataset selection and curation**: Knowing the source of data and making an initial assessment of the data

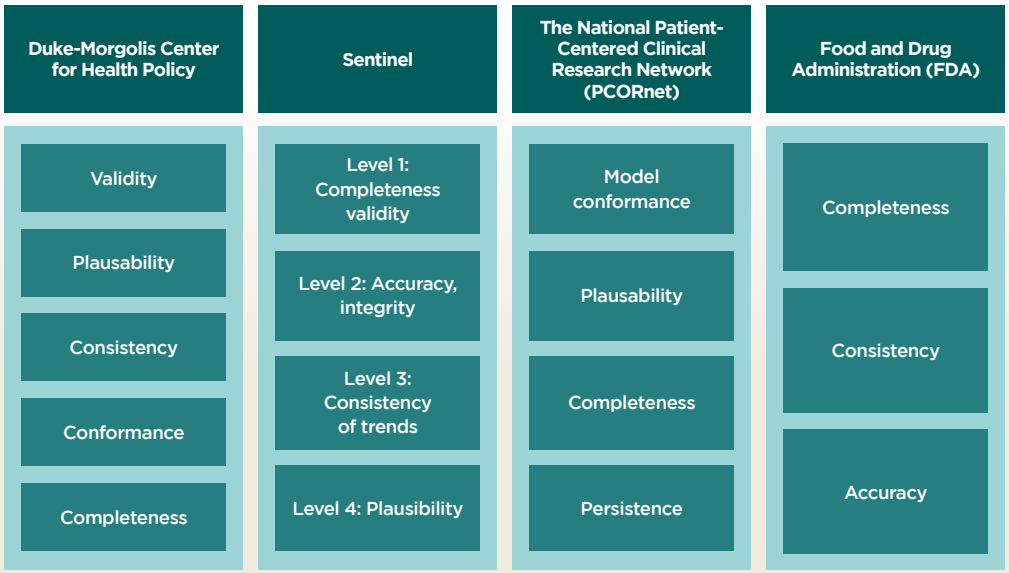
quality can help to determine the potential for selection and information bias. Selection bias results when the data used to produce the model are not fully representative of the actual data that the model may receive or of the environment in which the model will function. In addition to selection bias, measurement bias is another relevant issue that results when the data collection device causes the data to be systematically skewed in a particular direction. Consequently, developers should be aware of data quality limitations when attempting to curate and utilize these large-scale datasets. Moreover, developers and regulators need to know where the data originally came from and how the information was collected and curated. This is especially important when the datasets are from an open-source database where the original source and specifications of the dataset may not be available. When the origin of data is difficult to establish, it would be prudent for developers to assess the risks of using such data and manage them accordingly. Finally, even if datasets are collected from reliable sources, the mitigation of bias and assessment and mitigation of other risks to data robustness remain essential for a heterogeneous dataset.

- **Data usability**: It is essential to know whether the data used for the development of the algorithm was intended for that training, so developers need to convey their full understanding of the dataset and why it was suitable for their purpose. For instance, data from a third-party source may be representative data intended for training purposes (e.g. case studies in tertiary education) and may not be suitable for training an AI model intended to diagnose a disease or condition.
- **Data integrity**: Data integrity can be defined as "the completeness, consistency, and accuracy of data" (FDA, 2024). Lack of data integrity is an important issue. This can be best understood by how well extraction and transformation have been performed on the dataset. To maintain data integrity, data verification checks may be developed. Data verification checks are a key component of data quality assurance when utilizing real-world

data. Such checks should also be the first step in data preparation for any ML workflow.

- **Model training**: AI algorithms are usually trained on a separate dataset (known as the training dataset) and validated on a different dataset in order to reliably measure the performance of the algorithm. Training datasets should be well represented (e.g. by considering the prevalence of a disease/condition) to avoid "class imbalance". Medical record data is inherently biased, and therefore it is necessary to incorporate non-medical data such as the social determinants of health (Obermeyer et al., 2019). Furthermore, under-representation of important diagnostic features may limit the model's performance and cause bias. This can be avoided by ensuring that inclusion and exclusion criteria at both the patient level and the data input level do not create a selection bias. Furthermore, when ensuring that the datasets reflect the setting in which the model will be applied, a lack of diverse data (age, race, geographical areas) could limit the generalizability and accuracy of a developed AI system. This is demonstrated by a recent study from Stanford University (Shana, 2020), which showed that 71% of patient data from just three US states train most of the AI diagnostic tools used in the United States of America.

FIGURE 13 - EXAMPLES OF QUALITY CHECK PRINCIPLES



SOURCE: DUKE-MARGOLIS CENTER FOR HEALTH POLICY (2019).

- **Data labelling**: It is important to ensure consistent, reliable, and accurate labelling of datasets for testing in line with good practices. In cases where subjective reference standards are used, quality will be influenced by many factors, such as the independence and qualifications of the graders, the number of graders per label, whether the reference standard is validated to correlate with patient outcomes, and whether the reference standard follows published metrics.
- **Documentation and transparency**: The algorithm and data supporting it are often not available or are not well documented for all AI system stakeholders. This makes it difficult to assess the quality of the underlying data. Transparency and careful documentation are important not only with regard to the methodology used in collecting data, but also for the selection and modifications of datasets used for training, validation and testing. Good documentation is fundamental for

achieving transparency, which enables verification and traceability. Transparency of methods should ensure data quality. Beyond the CONSORT-AI and SPIRIT-AI reporting guidelines, checklists have been devised by the machine learning community to report representativeness, completeness and other data quality characteristics[4] (Gebru et al., 2021).

In addition, developers should consider deploying rigorous pre-release trials for AI systems to ensure that they will not amplify any of the issues discussed, such as biases and errors in the training data, algorithms, or other elements of system design. Furthermore, careful design or prompt troubleshooting can help identify data quality issues early. This could potentially prevent or mitigate possible resulting harm. Finally, to mitigate data quality issues that arise in healthcare data and the associated risks, stakeholders should continue working to create data ecosystems to facilitate the sharing of good-quality data sources.

The list in Table 3 summarizes the key data quality considerations for AI system safety and effectiveness.[5]

4 Find out more: https://datanutrition.org/

5 This list will be updated and harmonized with the work of the IMDRF.

TABLE 3 - GENERAL DATA QUALITY CONSIDERATIONS

CATEGORY	DATA QUALITY CONSIDERATION ITEM
Dataset	• Splitting • Selection volume and size • Selection bias • Individual variables' definitions in each dataset • Raw data versus "cleaned" data • Data wrangling and cleansing • Parameters and hyperparameters • Usability • Characterization • Labeling • Dependencies • Augmentation • Manipulation • Streaming • Interfaces • Integrity • Unique requirements • Data source
Data infrastructure	• Storage size • Storage format • Transformation medium
AI/ML model	• Data training • Tuning data • Verification set • Validation set • Testing • Development set • Static AI versus dynamic AI • Open AI versus closed AI
Governance management	• Liability • Data access • Risk management • Data protection • Privacy • Adoption education for clinical practice • Good practices • Standards (of care, governance, interoperability, etc.) • Scope of practice and AI model use • Technical checklist • Documentation • Transparency

SOURCE: PREPARED BY THE AUTHORS.

PRIVACY AND DATA PROTECTION

The WHO Global Strategy on Digital Health 2020-2025 classifies health data as sensitive personal data, or personally identifiable information, which requires a high standard of safety and security. Therefore, the strategy emphasizes the need for a strong legal and regulatory framework to protect the privacy, confidentiality, integrity, availability, and processing of personal health data. A responsive legal and regulatory framework can also address issues of cybersecurity, trust-building, accountability and governance, ethics, equity, capacity- building and literacy. This will help ensure that good-quality data are collected and subsequently shared to support the planning, commissioning, and transformation of services.

To develop and maintain adequate data security strategies, it is important for AI system developers, deployers, and manufacturers to understand the thickening web of privacy and data protection laws. This section discusses high-level considerations for privacy and data protection. For other ethical considerations, refer to the deliverable of the Working Group on Ethical Considerations on AI for Health[6] (WHO, 2021b).

Current landscape

As the demand for health-related data increases, the protection of privacy is creating a unique challenge for all stakeholders wishing to benefit from the many opportunities created by AI systems and technologies. One of the main reasons for this is that the high dimensionality of big data could make it difficult to apply anonymization and de-identification methods. Additionally, ensuring that large-scale datasets are secure from unauthorized access at each stage of the development process, collection, storage and management, transport, analysis, sharing, and destruction, is an important consideration.

Some 145 countries and regions have data protection regulations and privacy laws that regulate the collection, use, disclosure, and security of personal information (Greenleaf, 2021). There are many different definitions and interpretations

6 For a broader discussion of privacy and other ethical considerations for the use of AI, refer to the deliverable of the FG-AI4H's Working Group on Ethical Considerations on AI for Health and international, regional and national recommendations.

of "data protection" and "privacy". In some cases, data protection and privacy are used interchangeably. However, although these concepts are similar and often overlap, their meanings are different, and developers should be aware of the legal and ethical implications that result from these differences.

Laws and regulations that cover "the management of personal information" are typically grouped under "privacy policy" in the United States and under "protection policy" in the European Union (EU) and elsewhere. These laws are often complex and may have conflicting obligations. When developing an AI system for therapeutic development or healthcare applications, early in the development process the developers should consider gaining an understanding of applicable data protection regulations and privacy laws, including special regulatory provisions related to sensitive information such as genetic data. Developers should also consider national laws as well as regional ones. For instance, in the United States, although the Health Insurance Portability and Accountability Act (HIPAA) sets a baseline for protecting health data, states are empowered to enact stricter privacy laws (e.g. California's Consumer Privacy Act of 2018).

It is important to understand the jurisdictional scope of the various laws. For instance, because the scope of the GDPR is broad and its impact is significant, companies may want at least to evaluate the extent to which they are subject to it. Most privacy laws, including Singapore's Personal Data Protection Act, apply only to personal data processed within the country, whereas the GDPR[7] may apply to the personal data of EU citizens, regardless of the location where data are processed.[8] As a result, companies subject themselves to compliance obligations under the GDPR if they are located in the EU (including if any component of the organization is located in the EU), if they offer goods and services to individuals located in the EU, or if they monitor the behavior of persons located in the EU.

It is also important for developers to understand the varied legal contexts and requirements for privacy-related concepts

7 See also India's proposed Personal Data Protection Act.

8 Like the GDPR, the CCPA applies to natural persons who are California residents who are "domiciled in the state or who is outside the state for a temporary or transitory purpose." Cal. Code Regs. tit. 18, para. 17014.

such as "identifiable," "anonymous," and "consent". For example, Chapter 1 of the United Kingdom of Great Britain and Northern Ireland's draft anonymization, pseudonymization, and privacy-enhancing technologies guidance warns that referring to datasets as "anonymized" when they still may contain personal data in a pseudonymized form poses the risk of violating the United Kingdom of Great Britain and Northern Ireland's data protection law in the mistaken belief that the processing does not involve personal data (Information Commissioner's Office [ICO], 2021). Consent requirements also vary according to the jurisdiction. For instance, various jurisdictions may require "explicit consent," with heightened information requirements for the processing of health-related data (GDPR Article 9) (Regulation (EU) 2016/679). Therefore, developers may wish to consider the varied legal contexts when documenting how they address privacy-related concepts, including measures taken to meet consent requirements, and how they define anonymous or identifiable information.

In addition, certain jurisdictions have data protection regulatory frameworks that introduce reciprocity-based rules and place restrictions on the movement or transfer of data across borders. This may have a significant impact on the way in which data are processed and shared between countries. These provisions serve to curtail transnational data flows into and out of areas that are considered not to provide an "adequate" level of data protection.

Adequacy assessments may be required to determine whether a recipient country has thresholds of data protection laws and protections "essentially equivalent" or "substantially similar" to the jurisdiction from which the data were transferred. The GDPR, as a significant driver of emerging global data protection regimes, provides that the free transfer of personal data to third countries, non-European Union Member States, can primarily occur where the third country is considered by the EU Commission to have an "adequate" level of protection.[9]

9 Data flows have increasingly become an important part of global interconnection and AI development. Although the Schrems II case pertains to the EU-US position on data transfers, the wider implications inform global data transfers and the way in which they are to be compatible with GDPR requirements, including the validity of standard contractual clauses which depend on whether effective mechanisms are in place to ensure compliance with the level of protection required under the GDPR. *Data Protection Commissioner v. Facebook Ireland Limited, Maximillian Schrems* (Case C-311/18, "Schrems II").

As of May 2023, the EU Commission had recognized only 13 countries as providing adequate protection (EC, n.d.).

Developers should be aware of the nuances of the different jurisdictions' regulations and laws and should consider documenting their data protection practices accordingly. In general, companies should consider keeping abreast of new laws and requirements, leveraging governance, risk analysis, policies, training, and other strategies in a comprehensive and coherent way.

Documentation and transparency

Documentation and transparency are critical to facilitating trust with regard to privacy and data protection. Detailed privacy policy disclosures provide regulators with a benchmark by which to examine a company's handling of data. These disclosures should identify significant uses of personal information for algorithmic decisions. Depending on the jurisdiction, the disclosures may require the inclusion of other relevant information, e.g. the types and sources of health data collected and processed; the identities of the persons or organizations that determined the purpose or means of processing personal data; the identity of the person or organization which processed the data; the legal bases for processing the data; how the data were collected (including whether adequate notice was provided to the data subject and how consent requirements were met); and technical and organizational information on the storage of data, including security measures.

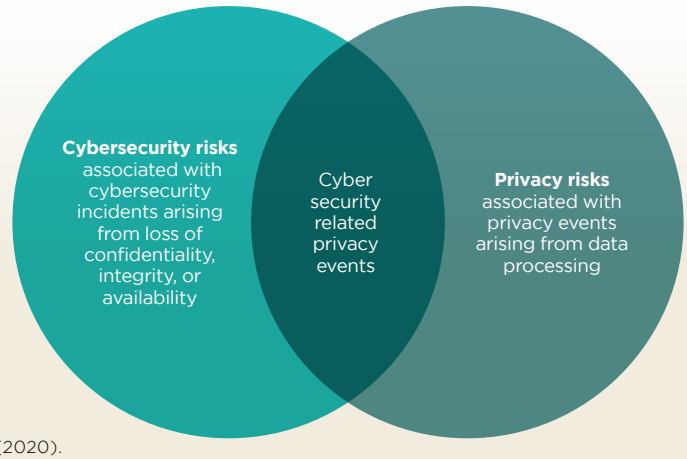
Developers must take privacy into account as they design and deploy AI systems. This includes designing, implementing, and documenting approaches and methods to ensure a quality *continuum* across the development phases to protect data privacy (Regulation (EU) 2016/679).[10] Privacy protections should not be limited only to addressing cybersecurity risks, especially since some privacy risks, such as harms to one's dignity

10 For example, a pillar of the data quality continuum in some jurisdictions, e.g., EU law, is the accountability principle. According to Art. 5 of the GDPR, data controllers shall abide by the five sets of principles enshrined in Art. 5(1), e.g., data minimization. Data controllers shall determine both technical and organizational measures to attain such ends (Art. 5(2)), throughout the entire cycle of data processing. Although not mentioned, the accountability principle is also at work in Art. 24(1), 25(1), and 32 of the regulation in regard to the responsibility of the controller, the principle of data protection by design (and by default), and security measures.

which may cause embarrassment or stigma, or more tangible harms such as discrimination, economic loss, or physical harm, (National Institute of Standards and Technology [NIST], 2020) can also arise by means unrelated to cybersecurity incidents. Therefore, when developing solutions to address risks, developers should have a general understanding of the different origins of cybersecurity and privacy risks and should develop their risk management practices accordingly (Figure 14).

A compliance program should consider risks and should develop privacy compliance priorities that take into account any specific potential harm as well as the enforcement environment. Developers may want to consider including in their documentation a description of the operations involved in the processing of personal data, a risk assessment, and the measures implemented to mitigate risks that take into account the interests of data subjects.

FIGURE 14 - NIST PRIVACY FRAMEWORK – CYBERSECURITY AND PRIVACY RISK RELATIONSHIP



SOURCE: NIST (2020).

Certain regulations outline prescriptive security requirements to address cybersecurity and privacy risks, such as the GDPR's data protection by design and default (GDPR Articles 25 and 32) (Regulation (EU) 2016/679) and India's proposed data privacy by design policy (Lei n. 22/2023), while others

include the duty to implement and maintain reasonable security practices and procedures appropriate to the risk.[11] Privacy frameworks often include privacy impact assessments, which are a widely used privacy management tool to proactively evaluate and mitigate privacy risks. Some jurisdictions, including the EU (GDPR Article 35) (Regulation (EU) 2016/679),[12] require companies to conduct these assessments.[13] Although the United States of America's law does not require privacy impact assessments, the NIST's privacy framework, of the US Department of Commerce, recommends that developers conduct them. According to NIST, "identifying if data processing could create problems for individuals, even when an organization may be fully compliant with applicable laws or regulations, can help with ethical decision-making in system, product, and service design or deployment" (NIST, 2020). This in turn can increase trust in the system.

Developers may also want to consider annotating their AI and having audit trails that explain what kinds of choices are made during the development process. Annotated notes provide "after the fact" transparency to outside parties and can help to explain the manner in which privacy was embedded, if applicable (West & Allen, 2020). These explanations and documentation should be available at different levels of detail, targeted at different audiences, such as regulators, managers, developers, operators, and users. The nature of the information and explanations required may differ, but all the assumptions, constraints, data sources, expected input and output, and major risks and limitations at each level should be clearly documented. In addition, an audit trail shows not only that controls have been applied but could also potentially show how damage was mitigated in the case of a data breach.

---

11 For example: CCPA § 1798.150(a)(1), South Africa's Protection of Personal Information Act of 2013; Israeli Privacy Protection Regulations (Data Security), 5777–2017 (implementing the Protection of Privacy Law, 5741–1981 of 1981); United Arab Emirates' Federal Law No. 2 of 2019; Kingdom of Saudi Arabia's E-Commerce Law of 2019 and its Implementing Rules.

12 "A data protection impact assessment shall be conducted if processing is likely to result in high risk to the rights and freedoms of the natural persons".

13 While risk assessments are quite common in information security standards and requirements, they are rarely seen in privacy rules in the United States of America. The GDPR, however, requires that an organization processing personal data must conduct a specific Data Privacy Impact Assessment or DPIA before beginning the processing.

Many jurisdictions enforce certain cybersecurity requirements or publish guidance on cybersecurity for consideration by developers of medical devices. Although an in-depth discussion of cybersecurity requirements is outside the scope of this subsection, it is important to understand the key role that cybersecurity plays in the protection of personal health information. Cybersecurity focuses on specific technical implementations needed to protect systems and networks against cyberattacks, which could compromise both the security of health-related systems and data as well as an individual's privacy, which could result in harm. To provide transparency about cybersecurity practices, developers may wish to consider documenting practices and approaches for data security, including policies that help protect the confidentiality, integrity, and availability of personal data throughout its lifecycle, such as appropriate encryption, access controls, logging methods, risk monitoring and methods of secure destruction. Developers may also consider documenting systems and approaches used to protect against data manipulation and adversarial attacks (NIST, 2018). For instance, blockchain-based technologies may be one mechanism for protecting data privacy, security, and integrity for AI in a traditionally fragmented health information systems ecosystem for national and regional contexts (Alsalamah et al., 2021).

AI regulatory sandboxes

The above regulatory challenges are recognized by regulatory authorities and policymakers across the world (Attrey et al., 2020). As a result, over 50 countries are currently experimenting with sandboxes in a wide range of high-technology sectors, notably in the financial sector but sandboxes have also gained popularity for health and legal services (Matiega & van de Pol, 2022). The regulatory sandbox approach has gained considerable traction as a means of helping regulators address the development and use of AI and other emerging technologies (Matiega & van de Pol, 2022). Regulatory sandboxes are generally regulatory tools that allow the flexibility to test innovative products or services with minimal regulatory requirements (Matiega & van de Pol, 2022). Consequently, regulatory sandboxes are considered an agile approach to

innovation and regulation and thus regulatory authorities are increasingly favoring them. In the EU, regulatory sandboxes have been proposed for testing surveillance solutions in the fight against the COVID-19 pandemic, and for establishing a framework for EU-wide data access. In relation to AI regulations specifically, the first AI regulatory sandbox pilot, presumably launched in 2023 by the Government of Spain,[14] aims to provide a guide to all EU Member States and the EC (CE, 2022). Although AI regulatory sandboxes raised a few concerns, they have the potential to bring many key benefits to AI system regulators, developers, manufacturers, and even patients (Matiega & van de Pol, 2022). This is because such AI regulatory sandboxes can: (a) help enable a better understanding of the AI systems during the development phase and before they are placed on the market; (b) facilitate the development of adequate enforcement policies and technical guidance that can mitigate risks and unintended consequences; and (c) foster AI innovation by establishing a controlled experimentation and testing environment for innovative AI technologies, products and services for new and potentially safer AI systems.

ENGAGEMENT AND COLLABORATION

Where applicable and appropriate, engagement and collaboration between developers, manufacturers, healthcare practitioners, patients, patient advocates, policymakers, regulatory bodies, and other stakeholders can improve the safety and quality of an AI system. Many regulatory bodies have adopted engagement and collaborative approaches in this area, and this section discusses the approaches of five of them: The United Kingdom of Great Britain and Northern Ireland's MHRA, the South African Health Products Regulatory Authority (SAHPRA), the European Commission, Singapore's HSA, and the US FDA. Table 4 lists examples of with whom, why, and how these regulators foster engagement

14 On November 9, 2023, the Ministry of Economic Affairs and Digital Transformation (Ministerio De Asuntos Económicos y Transformación Digital), of Spain, published the "Royal Decree 817/2023, of November 8, which establishes a controlled testing environment for the assessment of the conformity of the proposal for a Regulation of the European Parliament and of the Council establishing harmonized standards in the field of Artificial Intelligence." See Royal Decree 817/2023.

and collaboration. The examples are not meant to be comprehensive but instead are intended to highlight general approaches. Table 4 is followed by an analysis that discusses the similarities and differences in the approaches.

The subsection "Two successful instances of engagement" examines two examples of engagement and communication between regulators and AI developers resulting in positive clinical outcomes (CURATE.AI and IDentif.AI). The last subsections consider the practical implications for engagement and collaboration in resource-limited settings and recommend ways that regulatory bodies can initiate this process even in countries without past experience in engagement and collaboration. This is supplemented by several narratives: How to apply engagement tools (based on experience) and how to position the regulator as a partner in the context of accessible dialogue, and guidance and recommendations during the development process.

TABLE 4 - EXAMPLES OF REGULATORS' APPROACHES TO ENGAGEMENT AND COLLABORATION WITH STAKEHOLDERS REGARDING THE USE OF AI IN HEALTH CARE AND THERAPEUTIC DEVELOPMENT

1. (MHRA), United Kingdom of Great Britain and Northern Ireland	
With whom?	**Examples of stakeholders with whom the MHRA engages and collaborates:** • **Patients/patient advocates** • **Academia** • **Health-care professionals** e.g. providers in the National Health Service (NHS) and private healthcare providers. • **Industry** e.g. medical device and in vitro diagnostics industry, health technology industry. • **Domestic government partners** e.g. Department of Health and Social Care (DHSC), NHS England and Improvement, NICE, and Care Quality Commission (CQC).
Why?	**Examples of reasons why the MHRA engages and collaborates with stakeholders:** • Alert users to problems with medical devices and medicines. • Answer inquiries about roles in regulation or raise awareness of safety issues. • Seek feedback on the development of regulatory policy, managing adverse incidents, and risks. • Interface with the United Kingdom of Great Britain and Northern Ireland government and NHS, including stakeholders aligned to digital and AI-related activities.

How?	**Examples of ways in which the MHRA engages and collaborates with stakeholders:** • Central alerting system to the NHS and health-care providers or through professional groups. • Media, public, and other stakeholder inquiries via the MHRA customer service center, dedicated email inboxes, and press office. • Connecting with expert advisory groups, networks, and stakeholder groups on specific issues. • Consultation on engagement with patients and public. • Working-level meetings with national stakeholders, bilateral meetings with other parts of NHS, government, and international counterparts.

2. SAHPRA, South Africa	
With whom?	**Examples of stakeholders with whom the SAHPRA engages and collaborates:** • **Patients/patient advocates** • **Academia** • **Health-care professionals** • **Industry** (e.g. manufacturers/ distributors, trade associations). • **National government partners** (e.g. National Department of Health, National Department of Trade & Industry, South African National Accreditation Service).
Why?	**Examples of reasons why the SAHPRA engages and collaborates with stakeholders:** • Facilitate the approval of innovative AI systems. • South African National Accreditation System (SANAS) to ensure that the Conformity Assessment Body network is established in the country to certify the quality management system (QMS).
How?	**Examples of ways in which the SAHPRA engages and collaborates with stakeholders:** • The framework for engagement and collaboration has not yet been formalized. • Recommended that stakeholder engagement adopt the five-step engagement model developed by TGA.

3. EC, European Union	
With whom?	**Examples of stakeholders with whom the EC engages and collaborates:** • Patients/patient advocates • Academia • Health-care professionals

Why?	**Examples of reasons why the EC engages and collaborates with stakeholders:** • To "support the Commission in the development of actions for the digital transformation of health and care in the EU."
How?	**Examples of ways in which the EC engages and collaborates with stakeholders:** • By providing "advice and expertise to the Commission, particularly on topics set out in the communication[15] on enabling the digital transformation of health and care in the Digital Single Market, which was adopted in April 2018." In particular, such topics regard health data interoperability and record exchange formats, digital health services, data protection and privacy, AI, and "other cross-cutting elements linked to the digital transformation of health and care, such as financing and investment proposals and enabling technologies."

4. HSA, Singapore	
With whom?	**Examples of stakeholders with whom the HSA engages and collaborates:** • **Academia** (e.g. research institutions). • **Health-care professionals** • **Industry** (e.g. software and AI developers, trade associations). • **National government bodies**
Why?	**Examples of reasons why the HSA engages and collaborates with stakeholders:** • Early engagement and support to innovators to facilitate regulatory compliance, thus facilitating timely access to safe innovations for patients. • Actively consult on new policies and guidelines related to AI and software medical devices to receive and incorporate stakeholders' inputs and perspectives (Regulatory guidelines for software medical devices – a life cycle approach). • To work with other agencies responsible for the implementation and deployment of AI and software medical devices in the healthcare system to facilitate greater adoption of innovative technologies in the healthcare system.

15 Find out more: https://digital-strategy.ec.europa.eu/pt/node/3067

How?	**Examples of ways in which the HSA engages and collaborates with stakeholders:** • Rapid, streamlined engagement portals are available for several facets of product regulation[16] • Specific processes that can be straightforwardly addressed include the Medical Device Information Communication System (for application submissions for licenses, permits, registrations, etc.). • Online self-help tools to determine the product classification and risk classification for medical devices and simple forms to seek advice and confirmation from the HSA. • Medical Device Development Consultation: Online appointment booking system that allows innovators and developers to seek scientific and regulatory advice during the medical device development phase to facilitate regulatory compliance. • Online stakeholder consultation process for all new and revised policies and guidelines. • Regular focus group discussions and engagements with industry associations and companies.

5. FDA, United States of America	
With whom?	**Examples of stakeholders with whom the FDA engages and collaborates:** • **Patients/caregivers/patient advocates** • **Academia** (e.g. research institutions) • **Health-care professionals** • **Industry** (e.g. developers, device manufacturers, drug companies, trade associations). • **National government partners** (e.g. National Institutes of Health [NIH], Office of the National Coordinator for Health Information Technology [ONC], Federal Communications Commission [FCC]). • **Foreign government partners** • **International organizations** (e.g. IMDRF, ICH) • **Consumers/general public**
Why?	**Examples of reasons why the FDA engages and collaborates with stakeholders:** • Facilitate patient access to technologies that can benefit them in a timely manner. • Support novel, innovative medical product development through early interactions with stakeholders. • Provide timely feedback on FDA policies to reduce uncertainty. • Communicate to the public about AI/ML devices. • Receive feedback on policies, guidance, and discussion papers.

16 Find out more: https://www.hsa.gov.sg/e-services

How?	**Examples of ways in which the FDA engages and collaborates with stakeholders:** • Hold different types of pre-submission meetings to provide early feedback to sponsors. • Participate and lead international harmonization efforts (e.g. IMDRF, ICH). • Engage as members of public-private partnerships and collaborative communities. • Collaborate in pre-competitive space on regulatory science research to advance scientific community understanding. • Receive formal comments on policies and guidance through the Federal Register. • Hold workshops and other engagement events to obtain feedback from patients, industry, and other stakeholders.

SOURCE: PREPARED BY THE AUTHORS.

Discussion on strategies of profiled regulatory bodies

Table 4 shows the approaches of four national and one regional (in the case of the EC) regulatory bodies to foster engagement and collaboration. In the first category ("with whom?"), there are considerable similarities between these bodies. The shared targets for engagement and collaboration include health professionals (indicated by FDA, SAHPRA, MHRA, EC, and HSA), academia (FDA, SAHPRA, MHRA, EC, and HSA), industry (FDA, SAHPRA, MHRA, EC, and HSA), patients or patient advocates (FDA, SAHPRA, MHRA and EC), domestic government bodies (FDA, SAHPRA, and MHRA), media (national and trade press; FDA and MHRA), health providers (FDA and MHRA) and consumers (FDA and MHRA). Interestingly, the strategy paper by the US Department of Commerce's NIST also refers to academia and domestic government bodies as targets for engagement and collaboration.

In the second category ("why?"), SAHPRA notes the importance of communicating the benefits and intended use of devices, presumably to protect and promote public health (listed by the FDA and implied by MHRA). The FDA also stresses the importance of bilateral communication with stakeholders so that regulators are aware of developments in industry (or academia) and so that these stakeholders, in turn, are aware of developments in regulation. Similarly, MHRA indicates the importance of acquiring feedback about medical devices from stakeholders. This supports the objectives given

by both SAHPRA and the EC, namely, to facilitate the approval of innovative solutions and support the digital transformation of health and care. The HSA acknowledges the importance of early engagement with innovators and developers to provide greater clarity in regulatory requirements and improve transparency in regulatory processes.

For the third category ("how?"), the FDA lists steps that are taken to foster engagement (e.g. hosting workshops, producing digital and print material, and offering training modules or other types of education). MHRA also notes the importance of holding meetings with stakeholders (including domestic government institutes and international counterparts). HSA has introduced a pre-market consultation scheme to support innovation and device development by providing scientific and regulatory advice to enable regulatory compliance by software and AI developers who, unlike traditional medical device manufacturers, are not familiar with regulatory requirements[17] (Department of Health and Age Care, 2017).

Two successful instances of engagement

To understand the value of engagement and collaboration between regulatory bodies and stakeholders, two real-world examples (Case 1 and Case 2) are described. Clear avenues for engagement between regulators and AI developers play a major role in ensuring that rigorous evaluation and accelerated delivery of impactful modalities can be realized seamlessly. One aspect is in the area of interventional AI/digital medicine, which involves the application of software/devices (e.g. AI-based drug development and/or dosing platforms) and/ or the application of resulting drug compounds and/or combinations recommended by these platforms (Ho, 2020a; Ho, 2020b; Blasiak et al., 2020). In this context, integrating regulator accessibility with emerging innovation, sometimes in urgent circumstances, will ultimately result in life-saving outcomes. Importantly, these outcomes will not be confined to post-approval treatment but also to substantial patient benefit during the investigational stages of validation.

17 Find out more: https://www.iap2.org/

In **Case 1**, the developmental roadmap and validation of CURATE.AI and the foundational technology of IDentif.AI were discussed with the Medical Devices Branch (HSA, 2022) of the HSA in Singapore. This interactive session included an in-depth review of the key findings of the technology platforms, the process of implementing both platforms, emerging statistical analysis strategies to assess effectively the personalized medicine treatment outcomes and regulatory routes. A broader discussion on how clinical trial designs may evolve due to the emergence of AI was also conducted (Ho et al., 2020; Shah et al., 2019; Harrer et al., 2019). A clear pathway for subsequent inquiries was established, as multiple and frequent guidance requests were expected due to the nature of the trial designs that were envisioned. These included N-of-1 study designs for a broad range of indications designed for each patient. Specifically, these designs were personalized on the basis of (for example) the individualized dosage calibrations of the drug regimen (clinician-selected regimen), serial efficacy and toxicity measurements, efficacy-guided treatment protocols, and safety parameters. Subsequent submissions have included engagement with regulators for risk classifications associated with the device for each trial and subsequent discussion for submission of Special Access Routes (SAR) (HSA, 2019) for the potential rapid implementation of trials and for treatment purposes if needed. Rapid and informative responses and active engagement from HSA regulatory team members resulted in efficient turnaround times for trial initiation, which ultimately resulted in a positive outcome for a refractory oncology patient. A sustained track record of engagement with the regulatory community has played a key role in helping a clear process flow to be developed for downstream guidance requests.

The **Case 2** was developed in response to the COVID-19 pandemic. Specifically, a patient-derived live virus strain was harnessed for IDentif.AI-driven combination therapy optimization to serve as a clinical decision support system (CDSS). Unlike traditional AI-based approaches, this strategy did not use existing patient datasets. Instead, prospective experimentation was used alongside an AI-derived small data analytics strategy to pinpoint prospective data-backed

rankings of combinations for potential further clinical consideration and to potentially eliminate certain combinations from further clinical consideration. The foundational technology for IDentif.AI was previously discussed in detail with the HSA Medical Devices Branch, and additional IDentif.AI SARS-CoV-2 study information was provided in the context of clinical decision support, developing optimized combinations identified by IDentif.AI and with potential trials being designed with clinical partners. With regard to regulator engagement, the Medical Devices Branch of the HSA was contacted to provide device risk classification guidance for the submission of a Clinical Research Materials Notification (CRM-N) for study purposes. Obtaining a CRM-N is a required part of the submission of a clinical validation program because it stipulates the prerequisite of an initial assessment of device risk by the HSA (HSA, 2023). The submission portal and portal interaction were particularly straightforward to navigate and were integrated with a uniform access portal which was streamlined for efficient oversight and monitoring with regulatory bodies. This further demonstrates the straightforward process of interaction with the HSA. This case was an example of the critical importance of straightforward regulator accessibility and the profoundly positive impact that this can have on the advancement of promising technologies towards further clinical assessment and validation.

Recommended approaches for countries without past experience

For countries with limited experience in engagement and collaboration (and/or limited resources), it is important to establish: (a) what levels of engagement and collaboration are desired; (b) what steps can and should be taken to achieve those levels; and (c) what challenges are presented by the technology (e.g. AI explainability).

In many cases, it is desirable to adopt regulatory models that are adaptable, flexible, modular, and scalable in order to account for the uncertainties of innovation through appropriate oversight and coordination. These features fit not only the specific challenges of emerging technologies but also the regulatory approach of countries without past experience in this field or

with scarce economic resources. On the one hand, priorities should be scalable so that growing amounts of work can be suitably addressed by adding resources to the regulatory model. On the other hand, however, priorities should be determined in accordance with the modular adaptability of the steps and levels of engagement. In ecology, adaptability applies to the ability to cope with unexpected disturbances in the environment. In engineering, modularity refers to the interrelation of the separate parts of a software package or to the partitioning of the design to make it manageable. In multi-agent systems (MAS), it refers to the efficient usage of computational resources. We can profit from this notion to create adaptable policies that can be combined into regulatory systems for legal governance. The aim should be to address the uncertainties of innovation and to align with society's preferences on emerging innovation while allowing regulators to gain a growing understanding of technological challenges with increasing normative granularity (Pagallo et al., 2019).

Narrative on using engagement tools based on practical experience

For all countries, from those with limited experience in engagement and collaboration (and/or limited resources) to those at the other end of the spectrum, project, and program management tools can help organizations (including regulators) to structure and execute their engagement with stakeholders and users. No matter which tool is chosen, the key to valuable engagement is to invest time, energy, and thought into how best to engage stakeholders and then follow through on that engagement for the duration of a project or program. Engagement often fails if the investment is seen as a short-term rather than a long-term relationship.

The Australian Government's recommended five-step model for engagement (Department of Health and Age Care, 2017) is a good starting point for considering how a regulator could engage with developers of AI health products and services. In this model, engagement starts with thinking through the purpose of the engagement (based on what it is hoped to be achieved) and identifying the relevant stakeholders. When planning the different levels of engagement with stakeholders,

it is recommended to map out existing relationships and define the type of engagement and relationship that is needed with the stakeholder (and what type of relationship the stakeholder would be open to having). For instance, a digital health developer building an application (app) to support parents with children above a healthy weight may find that the primary health body concerned is an influential stakeholder that sets policies on managing children's weight. However, this is not a body with whom the developer of the app needs to engage regularly, so the developer may only "inform" the health body of the project. However, a developer will want to work with parents of children above a healthy weight to co-design the app and ensure that it fits their needs. It would, therefore, be important for the developer to "collaborate" with a representative group of parents and establish two-way or multi-way communication and shared learning and decision-making over the course of the project.

A similar approach for making sure that stakeholders are provided with the right information at the right time and are using optimal communication channels is outlined by one of the leading product development software companies (Atlassian, n.d.). Within the stakeholder communication "play", importance is placed on who the stakeholders are, the desired method of communication, and the frequency of communication. For instance, an internal government project developing a digital health product will have internal stakeholders (such as funders of the project and policy leaders) and external stakeholders (such as leading academics). The communications plan should outline how each stakeholder group will be addressed (e.g., email, face-to-face conversation, video call, and/ or social media) and how often there will be contact with the stakeholder group (e.g., daily, fortnightly, and/ or yearly) based on what the relationship with the stakeholder brings to the overall goals, such as information-sharing, co-design, and/or quality assurance. This plan can then be mapped out in a simple table (for which examples of headings might be: Method, audience/stakeholder, content to share, why, and frequency) for the whole development team to follow.

Narrative positioning the regulator as a partner in the development process

As demonstrated in Table 4 and discussed in the subsequent text, multiple regulatory bodies emphasize the importance of open (bilateral) communication with stakeholders so that regulators are aware of developments in AI-based technology and so that these stakeholders, in turn, are aware of changes in regulation. This is because AI-based technology is constantly changing, and regulation needs to be able to keep pace. The development, deployment, post-market surveillance and iteration of AI products and services in health care should therefore be an ongoing conversation between developers and regulators.

It is recommended that regulators look at AI-based technology in health care from a mindset of accessible engagement that potentially, when applicable, facilitates working alongside the developer to ensure compliance with regulatory requirements throughout the development and implementation process. An engagement mindset approach to regulation is about building trusting, collaborative relationships between developers and the regulatory body(s), and a two-way dialogue that enables developers to learn from regulators and vice/versa.

Furthermore, depending on a country's regulatory arrangements, one or more regulators may be responsible for AI-based health products and services. This means a developer often has to work with (and meet the standards of) more than one regulatory body. To ensure that this is a smooth and positive experience for AI developers, it is again recommended that regulators take a service approach. This means that a single, clearly marked pathway should be established and followed by an AI developer when ensuring the compliance of a product or service. Regulators need to collaborate with each other on issues such as clear messaging to developers and consistent levels of engagement with developers at the right point, and by sharing what they learn from different engagements with developers.

If a country wishes to take an accessible engagement approach to the regulation of AI products and services, co-regulation could be explored. As outlined by Clarke (2019), in a co-regulation approach regulators outlined a regulatory framework based on required compliance with the legislative act(s). The details of how this is applied in practice are

jointly developed by regulators and a representative sample of developers (Clarke, 2019). Similarly, when considering regulation from a service mindset, a co-regulatory approach, when appropriate and with any potential conflicts of interest properly managed, is about generating buy-in from developers by engaging them in the design and implementation of the regulatory process. The approach involves designing a regulatory process that reflects and acknowledges the needs of developers and not just those of the regulatory body and associated groups. Ultimately, however, regulators must remain fully independent of developers in order to make decisions that put the safety of the public first, as well as ensuring that public and private healthcare resources are used only for technologies that meet independently developed standards of quality, safety, and efficacy.

RECOMMENDATIONS FOR THE WAY FORWARD

Based on its work, the WG-RC recommends that stakeholders examine the key 18 considerations discussed in the previous section and summarized in Table 5 below as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development.

TABLE 5 - KEY RECOMMENDATIONS FOR REGULATORY CONSIDERATIONS ON AI FOR HEALTH BASED ON EACH OF THE SIX TOPIC AREAS

1. Documentation and transparency recommendations
1.1 Consider pre-specifying and documenting the intended medical purpose and development process, such as the selection and use of datasets, reference standards, parameters, metrics, deviations from original plans, and updates/changes during the phases of development. These should be considered in a manner that allows for the tracing of the development steps, as appropriate. 1.2 Consider a risk-based approach also for the level of documentation and record-keeping utilized for the development and validation of AI systems.
2. Risk management and AI systems development lifecycle approach recommendations
2.1 Consider a total product lifecycle approach throughout all phases in the life of a medical device: premarket development management, post-market management/surveillance, and change management. 2.2 Consider a risk management approach that addresses risks associated with AI systems, such as cybersecurity threats and vulnerabilities, underfitting, algorithmic bias, etc.

3. Intended use, and analytical and clinical validation recommendations

3.1 Consider providing transparent documentation of the intended use of the AI system. Details of the training dataset composition underpinning an AI system – including size, setting and population, input and output data, and demographic composition – should be transparently documented and provided to users.
3.2 Consider demonstrating performance beyond the training dataset through external, analytical validation in an independent dataset. This external validation dataset should be representative of the population and setting in which the AI system is intended to be deployed and transparent documentation of the external validation dataset and performance metrics should be provided. This external validation dataset should be appropriately independent of the dataset used for the development of the AI model during training and testing.
3.3 Consider a graded set of requirements for clinical validation based on risk. Randomized clinical trials are the gold standard for the evaluation of comparative clinical performance and could be appropriate for the highest-risk tools or where the highest standard of evidence is required. In other situations, consider prospective validation in a real-world deployment and implementation trial which includes a relevant comparator using accepted relevant groups.
3.4 Consider a period of more intense post-deployment monitoring through post-market management and market surveillance for high-risk AI systems.

4. Data quality recommendations

4.1 Consider whether available data are of sufficient quality to support the development of the AI system that can achieve the intended purpose.
4.2 Consider deploying rigorous pre-release evaluations for AI systems to ensure that they will not amplify any of the relevant issues, such as biases and errors.
4.3 Consider careful design or prompt troubleshooting to help early identification of data quality issues, which could potentially prevent or mitigate possible resulting harm.
4.4 Consider mitigating data quality issues that arise in healthcare data and the associated risks.
4.5 Consider working with other stakeholders to create data ecosystems that can facilitate the sharing of good-quality data sources.

5. Privacy and data protection recommendations

5.1 Consider privacy and data protection during the design and deployment of AI systems.
5.2 Consider gaining a good understanding of applicable data protection regulations and privacy laws early in the development process and ensure that the development process meets or exceeds such legal requirements.
5.3 Consider implementing a compliance program that addresses risks and develops privacy and cybersecurity practices and priorities that take into account potential harm and the enforcement environment.

6. Engagement and collaboration recommendations

6.1 Consider the development of accessible and informative platforms that facilitate engagement and collaboration, where applicable and appropriate, among key stakeholders of the AI innovation and deployment roadmap. and collaboration.
6.2 Consider streamlining the oversight process for AI regulation through engagement and collaboration in order potentially to accelerate practice-changing advances in AI.

SOURCE: PREPARED BY AUTHORS.

CONCLUSION

WHO recognizes the potential of AI in enhancing health outcomes by improving clinical trials, medical diagnosis, treatment, self-management of care, and person-centered care, as well as creating more evidence-based knowledge, skills, and competence for professionals to support health care. Furthermore, with the increasing availability of healthcare data and the rapid progress of analytics techniques, AI has the potential to transform the health sector to meet a variety of stakeholders' needs in health care and therapeutic development. For this reason, WHO and ITU are collaborating through the Focus Group on AI for Health (FG-AI4H) to facilitate the safe and appropriate development and use of AI systems in health care. The FG-AI4H's Working Group on Regulatory Considerations (WG-RC) on AI for Health consists of members representing multiple stakeholders, including regulatory bodies, policymakers, academia and industry, who explored regulatory and health technology assessment considerations and emerging "good practices" for the development and use of AI in health care and therapeutic development. This publication, which is based on the work of the WG-RC, is an overview of regulatory considerations on AI for health that covers the following six general topic areas: Documentation and transparency, Risk management and the AI Systems Development Lifecycle Approach, Intended use and analytical and clinical validation, Data quality, Privacy and data protection, and Engagement, and collaboration. This overview is not intended as guidance, regulation or policy. Rather, it is a list of key regulatory considerations and is a resource that can be considered by all relevant stakeholders in medical devices ecosystems, including developers who are exploring and developing AI systems, regulators who might be in the process of identifying approaches to manage and facilitate AI systems, manufacturers who design and develop AI-embedded medical devices, health practitioners who deploy and use such medical devices and AI systems, and those working in this area. The WG-RC recommends that stakeholders examine these key considerations and other potential ones as they continue to develop frameworks and best practices for the use of AI in health care and therapeutic development in relationship to the six topic areas.

The WG-RC recognizes that AI has been instrumental in rapidly advancing research in health care and therapeutic development. However, it also recognizes the evolving complexity of the AI landscape and the need for international collaboration to facilitate the safe and appropriate development and use of AI systems. Accordingly, international collaboration on AI regulations and standards is important for three reasons. First, sharing knowledge and best practices of evolving regulatory considerations could increase the speed of developing this regulatory landscape and reduce the gap between advancing technology and regulation. Second, international collaboration improves consistency in regulations, which is important as many tools are likely eventually to cross borders. Consistency of regulatory considerations for AI systems and technologies could improve standards and enable more rapid deployment. Third, international collaboration supports countries with less regulatory capacity by ensuring that these countries can also use tools with high standards, reducing the potential for disparity in the introduction of these tools. Eventually, the WG-RC understands that the AI landscape is rapidly evolving and that the considerations in this deliverable may need to be expanded as the technology and its uses develop. The working group recommends that stakeholders, including regulators developers, and manufacturers, continue to engage and that the community at large works towards shared understanding and mutual learning. In addition, established national and international groups, such as the IMDRF, GHWP, AMDF, and ICMRA, should continue to work on AI topics for potential regulatory convergence and harmonization.

REFERENCES

Alsalamah, S. A., Alsalamah, H. A., Nouh, T., & Alsalamah, S. A. (2021). HealthyBlockchain for global patients. *Computers, Materials & Continua*, *68*(2),2431–2449. https://www.researchgate.net/publication/351028022_HealthyBlockchain_for_Global_Patients

Atlassian. (n.d.). *Stakeholder communications*. https://www.atlassian.com/team-playbook/plays/ stakeholder-communications-plan

Attrey, A., Lesher, M., & Lomax, C. (2020). The role of sandboxes in promoting flexibility and innovation in the digital age. *Going Digital Toolkit Note, 2*. https://goingdigital.oecd.org/data/notes/No2_ToolkitNote_Sandboxes.pdf

Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y. T., Hamzah, H., Ho, J., Lee, X. Q., Hsu, W., Lee, M. L., Musonda, L., Chandran, M., Chipalo-Mutati, G., Muma, M., Tan, G. S. W., Sivaprasad, S., Menon, G., Wong, T. Y., & Ting, D. S. W. (2019). Artificial Intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study. *Lancet Digit Health*, *1*(1), e35–e44. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30004-4/fulltext

Blasiak, A., Lim, J. J., Seah, S. G. K., Kee, T., Remo, A., Chye, D. H., Wong, P. S., Hooi, L., Truong, A. T. L., Le, N., Chan, C. E. Z., Desai, R., Din, X., Hanson, B. J., Chow, E. K.-H., Ho, D. (2020).IDentif.AI: Rapidly optimizing combination therapy *design* against severe acute respiratory syndrome Coronavirus 2 (SARS-Cov-2) with digital drug development. *Bioengineering & Translational Medicine*, *6*(1), 1-16. https://aiche.onlinelibrary.wiley.com/doi/10.1002/ btm2.10196

Calderon-Ramirez, S., & Oala, L. (2021). More than meets the eye: Semi-supervised learning under non-IID data. In *The International Conference on Learning Representations*. https://arxiv.org/abs/2104.10223

Clarke, R. (2019). Regulatory alternatives for AI. *Computer Law & Security Review, 35*(4), 398–409.

Department of Health and Age Care. (2017). *Stakeholder Engagement Framework*. https://www.health.gov.au/resources/publications/stakeholder-engagement-framework

Duke-Margolis Center for Health Policy. (2019). *Determining real-world data's fitness for use and the role of reliability*. https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf

European Commission. (2022). *First regulatory sandbox on Artificial Intelligence presented*. https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented

European Commission. (n.d.). *Adequacy decisions: How the EU determines if a non-EU country has an adequate level of data protection*. https://ec.europa.eu/info/law/law-topic/data-protection/ international-dimension-data-protection/adequacy-decisions_en

Food and Drug Administration. (2019). *Proposed regulatory framework for modifications to Artificial Intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) - Discussion paper and request for feedback*. https://www.fda.gov/files/medical%20devices/ published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

Food and Drug Administration. (2021). *Artificial Intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) Action plan*. https://www.fda.gov/media/145022/download

Food and Drug Administration. (2024). *Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *ACM Communications*, *64*(12), 86–92. https://arxiv.org/abs/1803.09010

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. This law addresses the processing of personal data, including in digital media, by natural persons or legal entities, whether public or private, with the aim of protecting the fundamental rights of freedom and privacy, and the free development of the personality of the natural person. https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf

General Data Protection Regulation (GDPR). (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Greenleaf, G. (2021). Global tables of data privacy laws and bills. *Proceedings of the Privacy Laws & Business International Report*, 169, 6–19. http://dx.doi.org/10.2139/ssrn.3836261

Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial Intelligence for clinical trial design. *Trends in Pharmacological Sciences*, *40*(8), 577–591. https://www.sciencedirect.com/science/article/pii/S0165614719301300#:~:text=AI%20techniques%20have%20advanced%20to,to%20assist%20 human%20decision%2Dmakers.&text=We%20explain|%20how%20recent%20advances,towards%20increasing%20trial%20success%20rates

Health Sciences Authority. (2019). *Import and supply of unregistered medical devices by request of qualified practitioners*. https://www.hsa.gov.sg/ medical-devices/registration/special-access-routes/qualified-practitioner-request

Health Sciences Authority. (2023). *Complementary health products (CHP) classification tool*. https://www.hsa.gov.sg/CHP-classification-tool

Health Sciences Authority. (2022). *Regulatory guidelines for software medical devices – A lifecycle approach* (*online*). https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance- documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle- approach_r2-(2022-abr)-pub.pdf

Ho, D., Quake, S. R., McCabe, E. R. B., Chng, W. J., Chow, E. K., Ding, X., Gelb, B. D., Ginsburg, Hassenstab, J., Ho, C.-M., Mobley, W. C., Nolan, G. P., Rosen, S. T., Tan, P., Yen, Y., & Zarrimpar, A. (2020). Enabling technologies for personalized and precision medicine. *Trends in Biotechnology*, *38*(5), 497–518. https://www.cell.com/trends/biotechnology/fulltext/S0167-7799(19)30316-6

Ho, D. (2020a). Artificial Intelligence in cancer therapy. *Science 367*(6481), 982–3. https://science. sciencemag.org/content/367/6481/982

Ho, D. (2020b). Addressing COVID-19 drug development with Artificial Intelligence. Advanced Intelligent Systems, *2*(5). https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202000070

Information Commissioner's Office. (2021). *Introduction to anonymisation: Draft anonymisation, pseudonymisation, and privacy enhancing technologies guidance*. https://ico.org.uk/Media/About-the-ICO/Consultas/2619862/anonymisation-intro-and-first-chapter.pdf

International Medical Device Regulators Forum. (2013). *Software as a Medical Device (SaMD): Key definitions*. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf

International Medical Device Regulators Forum. (2014). *Software as a medical device: Possible framework for risk categorization and corresponding considerations*. https://www.imdrf. org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

International Medical Device Regulators Forum. (2016). *Software as a medical device (SaMD): Clinical evaluation*. http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-samd-ce.pdf

International Medical Device Regulators Forum. (2019). *Principles and practices for medical device cybersecurity*. http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-ppmdc.pdf

International Telecommunication Union. (2020). *Workshop on clinical evaluation of AI for health*. https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/ws/2010.aspx

International Telecommunication Union. (2021). *FG-AI4H Open Code Initiative – evaluation and reporting package*.

International Telecommunication Union. (2022). *Iniciativa de Código Aberto FG-AI4H (OCI)*. https://www.itu. int/en/ITU-T/focusgroups/ai4h/Pages/opencode.aspx

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K.; & Grupo de Trabalho Standar Protocol Items: Recommendations for International Trials for AI e Consolidates Standars of Reporting Trials for AI. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine, 26*,1364–1374. https://www.nature.com/articles/s41591-020-1034-x#citeas

Macdonald, J., März, M., Oala, L., & Samek, W. (2021). Interval neural networks as instability detectors for image reconstructions. In: Palm, C., Deserno, T. M., Handels, H., Maier, A., Maier-Hein, K., Tolxdorff, T. (Eds.), *Bildverarbeitung für die Medizin*. Informatik aktuell (Image processing for medicine. IT update). Springer Vieweg, Wiesbaden. https://doi.org/10.1007/978-3-658-33198-6_79

Madiega, T., & van de Pol, A. L. (2022). Artificial Intelligence act and regulatory sandboxes. *European Parliamentary Research Service*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf

National Institute for Health and Care Excellence. (2019). *Evidence standards framework for digital health technologies*. https://www.nice.org.uk/Media/Default/About/what-we-do/our- programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

National Institute of Standards and Technology. (2018). *Framework for improving critical infrastructure cybersecurity*. https://www.nist.gov/cyberframework

National Institute of Standards and Technology. (2020). *NIST privacy framework: A tool for improving privacy through enterprise risk management*. https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf

National Health Service. (2020). *A buyer's guide to AI in health and care*. https://www.nhsx.nhs.uk/ai-lab/explore- all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/

Notification of the Department of Medical Devices and Pharmaceutical Products No. 14, issued on August 31, 2021. (2012). Notification No. 0831-14, dated August 31, 2020 (Chinese). Handling of requests for PACMP confirmation for medical devices, PSEHB/SD (in Japanese). Tokyo: Ministry of Health, Labour and Welfare; 2020. https://www.mhlw.go.jp/content/11120000/000665757.pdf

Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A. W., Calderon-Ramirez, S., Li, D. X., Nobis, G., Alvarado, E. A. M., Jaramillo-Gutierrez, G., Matek, C., Shroff, A., Kherif, F., Sanguinetti, B., & Wiegand, T. (2020). ML4H Auditing: From paper to practice. *Proceedings of Machine Learning for Health (ML4H) NeurIPS Workshop*, 136, 280-317. https://proceedings.mlr.press/v136/oala20a.html

Oala, L., Johner, C., Goldschmidt, P. G., & Balachandran, P. (2022). Good Practices for Health Applications of Machine Learning: Considerations for Manufacturers and Regulators. *Proceedings of the Focus Group on Artificial Intelligence for Health*. https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-AI4H-2022-2-PDF-E.pdf

Oala, L., Heiß, C., Macdonald, J., März, M., Kutyniok, G., & Samek, W. (2021). Detecting failure modes in image reconstructions with interval neural network uncertainty. *International Journal of Computer Assisted Radiology and Surgery*, *16*, 2089–2097. https://link.springer.com/article/10.1007/s11548-021-02482-2

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://www.science.org/doi/10.1126/science.aax2342

Pagallo, U., Casanovas, P., & Madelin, R. (2019). The middle-out approach: Assessing models of legal governance in data protection, Artificial Intelligence, and the Web of Data. *The Theory and Practice of Legislation, 7*(1),1–25.

Panesar, A. (2019). *Machine learning and AI for healthcare: Big data for improved health outcomes*. Apress.

Royal Decree 817, of November 8, 2023. (2023). Establishes a controlled testing environment for the conformity assessment of the proposed Regulation of the European Parliament and the Council laying down harmonized rules on Artificial Intelligence. *Official State Bulletin*. https://www.boe.es/boe/dias/2023/11/09/pdfs/BOE-A-2023-22767.pdf

Rivera, S. C., Liu, X., Chan, A., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving Artificial Intelligence: The SPIRIT-AI extension. *BMJ*, 1351-1363.

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., & Strand, F. (2020). External evaluation of 3 commercial Artificial Intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncology, 6*(10), 1581–1588. https://pubmed.ncbi.nlm.nih.gov/32852536/

Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine, 3*(1), 1–4.

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., & Schork, N. (2019) Artificial Intelligence and machine learning in clinical development: A translational perspective. *npj, 2*(69), 1-5. www.nature.com/ artigos/ s41746-019-0148-3

Shana, L. (2020). The geographic bias in medical AI tools. Ethics and Justice, Healthcare, Machine Learning. *Stanford University Human-Centered Artificial Intelligence*. https://hai.stanford.edu/news/geographic-bias-medical-ai-tools

Topol, E. J. (2020). Welcoming new guidelines for AI clinical research. *Nature Medicine, 26*,1318–1320. https://www.nature.com/articles/s41591-020-1042-x#:~:text=A%20new%20era%20needs%20new,in%20conjunction%20with%20the%20algorithm.

Verks, B., & Oala, L. (2020)Data and Artificial Intelligence assessment methods (DAISAM) Audit Reporting Template. In: *Proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health*.

West, D. M., & Allen, J. R. (2020). *Turning point: Policymaking in the era of artificial intelligence*. Brookings Institution Press.

Willis, K., Oala, L. (2021). Adaptação de domínio post-hoc via homogeneização guiada de dados. *arXiv*, https://arxiv.org/abs/2104.03624

World Health Organization. (2020). *Guidance for post-market surveillance and market surveillance of medical devices, including in vitro diagnostics*. https://iris.who.int/handle/10665/337551

World Health Organization. (2021a). *Global Strategy on Digital Health 2020–2025*. https://iris.who.int/handle/10665/344249

World Health Organization. (2021b). *Ethics and governance of Artificial Intelligence for health: WHO guidance*. https://apps.who.int/iris/handle/10665/341996

World Intellectual Property Organisation. (n.d.). *Artificial Intelligence and intellectual property policy*. https://www.wipo.int/about-ip/en/artificial_intelligence/policy.html

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., & Zou, J. (2021). How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nature Medicine, 27*, 582–584. https://www.nature.com/articles/s41591-021-01312-x

Transparency and explainability: Prospects for the regulation of Artificial Intelligence in healthcare in Brazil

Daniel A. Dourado[1] *and Fernando Aith*[2]

1 Psychiatrist and health lawyer, with a master's and Ph.D. in Preventive Medicine/Public Health from the University of São Paulo's School of Medicine (FMUSP), he is also an associate researcher and professor at the Center for Research in Health Law at the University of São Paulo (Cepedisa/USP).
2 Lawyer; full professor in the Department of Policy, Management, and Health at the School of Public Health of USP (HSP/FSP); scientific coordinator of the Cepedisa at USP; editor-in-chief of the USP's Health Law Journal.

INTRODUCTION

The healthcare sector is undergoing a transformation driven by Artificial Intelligence (AI), which has the potential to revolutionize healthcare practices and services worldwide. The digital era has introduced innovative technologies capable of substantially changing the structure of health systems by opening up a range of unprecedented opportunities to improve the quality of individual and population care, expand access, reduce costs, and explore new frontiers in the prevention, diagnosis, and treatment of diseases (Matheny et al., 2022). Although the integration of AI in healthcare has progressed at a slower pace compared to other sectors, there is growing consensus on its inevitable integration into different domains of medicine and public health (Sahni & Carrus, 2023). For this technology to be used in a safe, transparent, responsible, and fair way, the development of specific frameworks for AI in healthcare is considered essential (World Health Organization [WHO], 2021).

In this context, the regulation of AI in healthcare assumes a central importance. AI systems need to be verified for their quality and safety, recognizing that healthcare actions and services, traditionally performed by people, have been significantly influenced and executed by automated systems. Therefore, while these new technologies adoption must be encouraged, it is essential to establish a regulatory structure capable of ensuring that their use is always for the benefit of human beings.

The tools developed for healthcare based on machine learning models are recognized as representing the biggest challenges in this scenario, due to their ability to learn from real-world experiences and adapt to improve their performance continually (Bates, 2023). Unlike the objects of traditional healthcare regulation, such as medicines and medical devices, AI systems can constantly change, even after they have been implemented (Gottlieb & Silvis, 2023).

An agenda for regulating AI in healthcare was implemented at the global level in the first half of the 2020s, with contributions coming from various entities such as the WHO, the International Medical Device Regulators Forum (IMDRF), the International

Organization for Standardization (ISO), the Organisation for Economic Co-operation and Development (OECD), and others. Attempts have been made to define the general principles that can be applied to AI's regulation in healthcare in different contexts: Low-, middle- and high-income countries, public and private sectors, governments, and organizations. The first relevant regulatory approaches to AI in healthcare relate to the dimensions of healthcare data governance and the safety and effectiveness of AI-based medical devices. The very rapid evolution of this overview has made this an increasingly complex activity (WHO, 2021, 2023, 2024).

The purpose of this article is to analyze another fundamental dimension in the regulation of AI in healthcare: Transparency. In doing so we will briefly explain the general concept of AI, look at its applications in health and medicine, present the ethical principles that serve as the basis for regulating AI in healthcare, and highlight the importance of transparency. We will then explore the interpretability and explainability concepts of AI systems, which are often associated with transparency's normative expressions. We will address the right to receive an explanation of the automated decisions that are taken as a legal mechanism of transparency, examining the current stage of the discussion of this matter in both international and Brazilian contexts. Finally, we will analyze the limits of current explanation techniques in AI in order to propose possible regulatory strategies based on the elements we identify.

APPLICATIONS OF AI IN HEALTHCARE AND MEDICINE

Since AI first emerged as a field of study, the healthcare sector has become one of the most suitable for its application. The first support systems for clinical activity were designed and developed by pioneering researchers in the 1950s. Different programs were created in the 1970s that aimed to simulate specialized human reasoning, the objective being to help doctors formulate diagnostic hypotheses in complex cases, interpret clinical exams, and select the appropriate treatments. These programs, which were called "clinical decision support systems" (CDSS), were significant applications of the so-called "symbolic AI," which was the dominant paradigm at the time

and was used during the 1980s and 1990s (Sutton et al., 2020). These rule-based systems, however, showed the limitations of this technological approach, particularly their high maintenance costs and the need for constant updates, which required frequent reviews by specialists. Furthermore, the performance of these systems was also constrained by the accuracy of previously existing medical knowledge. (Yu et al., 2018).

In this scenario, the emergence of the machine learning field has been received with great enthusiasm in health and medicine. The technique is a subtype of AI that offers tools for developing systems that can identify previously unknown patterns in datasets, without the need to pre-define the decision rules for each specific task. The growing availability of large volumes of data in healthcare, combined with the exponential increase in computing capacity, has driven the increasing expectation that AI will be substantially incorporated into the sector, an expectation that has intensified in particular since the 2000s.(Rajkomar et al., 2019). The recent expansion in interest in AI in healthcare, particularly since the mid-2010s, is also due to the successful implementation of techniques in several domains that are taken from a subtype of machine learning known as deep learning (Hinton, 2018).

The application of AI in health and medicine has quickly grown. It started in areas that deal with identifying patterns in images, such as radiology, pathology, and dermatology, and has expanded beyond computer vision to encompass areas such as natural language processing for analyzing data in electronic health records, the analysis of genetic information in precision medicine, and reinforcement learning in robot-assisted surgery. These techniques have performed in a promising way and made useful and accurate predictions in different clinical scenarios (Topol, 2019). They can also be very useful at the collective level in decision-making in public health and planning the allocation of human and financial resources. An important example of this is the use of AI tools for identifying and tracking infectious disease outbreaks and monitoring mitigation strategies. (Brownstein et al., 2023). New AI tools also have the potential to significantly reduce the time and costs associated with traditional drug discovery methods (Nature, 2023).

In recent years, the healthcare sector has seen the emergence of a new frontier with the rapid rise of foundation models since 2020, particularly large language models (LLM). The applications of foundation models in healthcare are vast and have the potential to transform the area. The models developed in this paradigm have demonstrated their ability to interpret different types of medical information, such as images, electronic medical records, laboratory test results, genomic data, and medical texts. These models can also provide different results, such as explanations in everyday language, recommendations, and annotations: They can even interact with humans (Moor et al., 2023). The expansion of these models to become multimodal – i.e., capable of interpreting not only text, but also images, audio, and video – exponentially increases their possibilities for application in individual and collective healthcare situations (Acosta et al., 2022).

ETHICAL PRINCIPLES FOR REGULATING AI IN HEALTHCARE

The regulation of AI has become a global priority in the 2020s (G7 Hiroshima Summit, 2023). The regulatory approach began its structuration through codes of conduct and non-binding guidelines (soft law) produced by government entities, expert councils set up to advise public bodies, research institutes, and private companies. The main foundations derive from the interdisciplinary field of ethics in AI, which addresses the moral, legal, and social implications of this technology with the aim of guiding its development and use in harmony with human values and social norms (Dubber et al., 2020). In recent years, there has been global convergence around five ethical principles: (a) transparency; (b) justice and equity; (c) non-maleficence; (d) responsibility; and (e) privacy. There is still significant divergence, however, regarding the meaning of these principles, how they should be interpreted, and the path to implementing them (Jobin et al., 2019).

The debate about the importance of defining specific ethical principles for regulating AI in healthcare began in the second half of the 2010s when the first evidence emerged that the field of machine learning would be highly promising and revolutionary in healthcare (International Bioethics Committee, 2017). In

2021, the WHO published its inaugural guide, with guidelines on the ethics and governance of AI in health, the aim being that this would have a global reach (WHO, 2021). This document consolidates the first basic principles considered consensual in the field of AI ethics that apply specifically to health: (a) protecting autonomy; (b) promoting human well-being, human security, and the public interest; (c) ensuring transparency, explainability and intelligibility; (d) promoting responsibility and accountability; (e) ensuring inclusion and equity; and (f) promoting responsive and sustainable AI.

The WHO has led the process of defining a global digital health strategy for supporting national health systems, which includes the preparation of an AI governance and regulation structure. In this context, the WHO's principles for AI ethics in health are intended to guide developers, users, and regulators when it comes to developing, implementing, and constantly evaluating the AI technologies that are being used in healthcare. The guidelines are based on a combination of the basic bioethical principles: Autonomy, non-maleficence, beneficence, and justice (Beauchamp & Childress, 1979), and the currently recognized general principles of AI ethics: privacy, transparency, and accountability (Floridi et al., 2018). Among other measures to be taken, the WHO intends to work in a coordinated manner with intergovernmental entities to identify and formulate laws and policies. It will also consider an initiative to draft model legislation that governments that intend to create their own regulations for AI in healthcare can use as a reference (WHO, 2021, 2023).

Transparency is the ethical principle most frequently observed in codes of conduct that define general guidelines for the use of AI (Jobin et al., 2019). In the context of AI ethics applied to healthcare, transparency is linked to the recommendation that systems be clear and understandable for developers, regulators, and users, including healthcare professionals and patients. To this end, it is essential that pertinent information about AI systems be documented before they are implemented and continue to be disclosed on a regular basis after being approved for use. It is also essential in this context to facilitate public consultation and an understanding of how AI models work in the real world (WHO, 2021). Additionally, it is expected

that these technologies will be explainable in accordance with the ability to understand those to whom the explanation is directed, and will clarify the functioning and decision conditions of the algorithms for healthcare professionals, patients, and other users of the systems (Watson et al., 2019).

TRANSPARENCY AS THE BASIS FOR REGULATING AI IN HEALTHCARE

Recognition of transparency as a regulatory dimension for the application of AI in healthcare stems from the widespread acceptance of this fundamental ethical principle. The broad concept of transparency implies making all information available that justifies the decisions to all the parties involved, based on the results generated by AI systems. This aspect is considered essential for establishing society's trust in AI technologies (European Commission, 2019; Floridi et al., 2018). Broadly speaking, therefore, transparency refers to explaining the institutional context in which AI systems are designed, implemented, and managed, and that focuses on the comprehensive understanding of those people and organizations that are responsible for developing, using, and regulating AI.

Transparency can be promoted by mechanisms such as the adoption of standard documentation on the creation, training, and implementation of AI systems and models, as well as by way of processes for evaluating the impact of these systems in different application contexts (Mittelstadt, 2022). Accurate and extensive documentation should be the primary mechanism for ensuring transparency in AI applications in healthcare; thus, the adequate recording of clear and detailed information about methods, resources, and the decisions taken throughout the entire life cycle of AI systems becomes an essential regulatory requirement. Regulators must have access to adequate documentation that covers everything from design, development, training, and the validation of models to implementation and the post-implementation period (WHO, 2023). This means demanding accurate information about the assumptions and limitations of AI systems to include operating protocols and data selection, processing and labeling methods, and the conditions for developing and validating machine learning models.

The term "transparency" is also widely used in the field of AI, particularly in machine learning, in more specific ways. The notion of transparency can be linked to an understanding of how algorithms work, without necessarily analyzing the training data or individual predictions of a model (OECD, 2019), a perspective that is usually referred to as "algorithmic transparency" (Association for Computing Machinery US [USACM] Public Policy Council, 2017). Alternatively, transparency can be understood as access to the essential elements for understanding how a model makes decisions, which implies not only knowing the algorithm but also the trained model and the data used in training it. To achieve this goal, it is essential to have a broad understanding of the model's features and of each of the components it has learned, such as weights and parameters (Lipton, 2018; Molnar, 2022). This conception of transparency is related to concepts of the interpretability and explainability of AI systems. Understanding these concepts is crucial for defining how such aspects should be considered in regulating AI in healthcare.

INTERPRETABLE AI AND EXPLAINABLE AI

In the dimension of transparency in AI, the terms "interpretability" and "explainability" are closely interconnected and often overlapping, because there are still no widely agreed definitions or limits in this field (Mittelstadt, 2022). A strong definition of interpretability points to the degree to which the cause of a decision made by an AI model can be understood by a human observer (Miller, 2019). Thus, a model is considered fully interpretable when a human being is able to discern the complete set of causes that gave rise to a given result; in contrast, an opaque or "black box" model is characterized by a lack of concrete information about how or why a specific result was obtained from the inputs (Burrell, 2016). Therefore, transparency can be considered at both the level of the model as a whole and the level of its components. An interpretable model can be a simple model, in which a human is able to connect the input data with the parameters for making calculations and generating a prediction, or a decomposed model, such that each part (input, parameter, calculation) allows for a more or less intuitive understanding (Lipton, 2018).

The idea of "explanation" is often associated with the way in which a human observer can gain understanding (Miller, 2019). An explanation is understood as a post hoc interpretation, by which information is derived from algorithmic results after decisions or predictions have been made, a process that can occur even in opaque models (Lipton, 2018). From this perspective, explainability can be considered a characteristic that is independent of the intrinsic interpretability of the model, since the explanation can be obtained by specific techniques in models that are not naturally interpretable. Consequently, the term "explainability" has been commonly used to refer to intrinsically interpretable models and to explanations obtained for models that are considered to be black boxes (Amann et al., 2022). Despite a frequent overlapping of meanings, using this framework it is possible to adopt definitions of interpretable AI and explainable AI that are particularly useful for application in the healthcare field (Babic et al., 2021).

Interpretable AI has to do with machine learning systems that are based on models that can be interpreted by humans. Although not everyone will necessarily understand these models immediately, people with knowledge in the field should be able to interpret them. It is important to note that, in practice, achieving global interpretability of models is very challenging since a model becomes inconceivable to the human mind after a certain number of characteristics: Imagining a space with more than three dimensions, for example. Therefore, an interpretable AI approach requires the models' use with a limited number of parameters and weights, so they are intelligible. It must also be based on "white-box" algorithms, such as linear functions, in which parameters correspond to weights that relate inputs and outputs, or decision trees, which produce intuitive maps based on clearly established and understandable rules (Molnar, 2022).

Explainable AI represents a fundamentally different approach since it involves black-box models, which are intrinsically incomprehensible to humans. Instead of trying to directly understand these opaque models, this approach adopts an alternative method: A second algorithm is trained to replicate the predictions of the opaque model, that is, an explanatory model is created that searches for an interpretable

function that most closely resembles the outputs of the black-box model. This surrogate model is then used to provide post hoc explanations for the original model's predictions, although it is incapable of accurately making actual predictions because it needs to simplify the number of characteristics to make it understandable. Explanations can highlight which attributes of the input data in the opaque model are most relevant to a specific prediction or create an easy-to-understand linear model whose results are similar to those of the original model. In short, explainable AI focuses on the task of finding a "white-box" model that can explain the predictions made by a black-box model (Babic et al., 2021).

BASES OF THE RIGHT TO EXPLANATION

The concept of the right to explanation of automated decisions involves recognizing the need to guarantee that everyone has the right to know how the AI-based decisions that affect their lives are made. This concept has been mainly consolidated from discussions that took place when preparing the European General Data Protection Regulation (GDPR), which was approved in 2016 and came into force in May 2018 (Regulation (EU) No. 2016/679).

According to European regulation, data subjects have the right to be informed when automated decisions significantly affect them or generate effects in their legal sphere. They also have the right to receive "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing" (Articles 13, 14, and 15 of the GDPR). Every data subject also has the "right not to be subject to a decision based solely on automated processing, including profiling," unless the decision is necessary for entering into or performing a contract or is based on the explicit consent of the subject. In these cases, the GDPR grants the data subject the right to request human intervention and to contest the automated decision (Article 22). The purpose of these safeguards is outlined in "Recital 71" of the GDPR, which provides interpretative guidance on these provisions. The data subject must have "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment

and to challenge the decision." In addition to receiving an understandable explanation, the right to the opportunity to be heard, question and request a review of the automated decision is established, a principle called "algorithmic due process"(Kaminski, 2019).

Since the GDPR was enacted, and both before and after it came into force, there have been debates about the existence and scope of the right to explanation concerning automated decisions (Bygrave, 2020). This issue fundamentally arises from the acceptance that the complex mathematical representation of machine learning models is, for the most part, incomprehensible to humans, especially since the increased use of black-box algorithms. The debate is generally divided into two perspectives. On the one hand, there are those who defend the viability and scope of the right to explanation only regarding the general functionality of the system, rather than the individual decisions made or specific circumstances (Wachter et al., 2017). On the other hand, there is an understanding that the explanation must also cover specific decisions, with transparency limited only by the intrinsic opacity of the algorithms, thus allowing the data subject to exercise their rights in accordance with the GDPR and compliance with the principles and laws of human rights (Selbst & Powles, 2017).

An alternative understanding is that the GDPR establishes a system of "qualified transparency" regarding algorithmic decision-making. This interpretation holds that the law defines targeted rules with different degrees of depth and scope aimed at different recipients, requiring one type of information to individuals and another type to experts and regulators (Kaminski, 2019).

The importance of the right to explanation in the healthcare area is to give human beings the possibility of understanding the logic of automated decisions that have an impact on their conduct in the care they receive. This concern must be increasingly present in various clinical situations. Some AI systems, for example, are currently able to define the criteria for organ transplants, such as allocation, matching donor and recipient, and predicting the survival times of transplant patients (Khorsandi et al., 2021). Systems of this type will possibly start being used in practice and there will

be differences in the order of transplant queues compared to those defined by clinical criteria made only by humans. Therefore, the right to explanation is related to human dignity, since decisions of this nature could not be made based exclusively on black-box systems.

RIGHT TO EXPLANATION AND TRANSPARENCY OF AI IN BRAZILIAN LEGISLATION

The debate that started with the European paradigm has been influencing other jurisdictions, such as Brazil's, where the General Personal Data Protection Law (LGPD) was approved in 2018 (Law No. 13.709/2018), which has been in force since August 2020 (the rules on administrative sanctions came into force in August 2021). The LGPD draws its inspiration from the GDPR and mirrors many of the institutes created by the European standard (Aith & Dallari, 2022). So, although the LGPD does not specifically deal with the regulation of AI (the terms "Artificial Intelligence" and "algorithm" do not even appear in the text of the law), it introduces the legal bases of the right to explanation and to a review of automated decisions into the Brazilian legal system.

The right to review automated decisions is explicitly defined in Art. 20 of the LGPD, which grants the subject the right to "request a review of decisions taken based solely on the automated processing of personal data that affect their interests" (Law No. 13.709/2018). Unlike the GDPR, the Brazilian law does not provide for the right not to be subjected to exclusively automated decision-making, or to obtain human intervention in the event of a review (Regulation (EU) No. 2016/679).[3] The right to explanation does not appear in the text in the Brazilian law (neither does it in the GDPR) but arises from the systematic interpretation of the LGPD itself in conjunction with constitutional provisions and consumer protection legislation (Monteiro et al., 2021). Brazilian law guarantees

3 In the original wording that was approved by the National Congress (August 2018), Article 20 of the LGPD provided for the right of the subject to request a review of automated decisions "by a natural person." However, this provision was altered by Provisional Decree No. 869 of December 2018, which removed the possibility of obtaining human intervention. This alteration was maintained when the provisional decree became law (Law No. 13.853/2019), which has defined the wording of the LGPD until the present date.

everyone affected by automated decisions the right to obtain clear and adequate information regarding the criteria and the procedures used. This expression of transparency can only be guaranteed by some form of explanation.

The rights associated with explanation and a review of the decisions taken by AI systems are necessarily linked and need to be understood together. As should happen in other countries in which the European model is used as a basis, these rights in Brazil still require regulation and future doctrinal and jurisprudential preparation. Several of the LGPD's provisions protect commercial and industrial secrets, so this consideration must be set out in infra-legal regulation, and even in the analysis of specific cases. The LGPD's protection of business secrets can be seen as a way of promoting a business model based on algorithms, still, it must necessarily be balanced with the right to an explanation of automated decisions in order to observe the ethical principles of using AI in harmony with human rights (Dourado & Aith, 2022).

The legislative process for regulating AI in Brazil is currently ongoing. Different proposals have already been suggested due to rapid changes in the field, and relevant changes have been made to the text in relation to the original version of the bill, which is now more mature (Bill No. 2338/2023).[4] It is worth noting that in ongoing debates in Brazil, the idea of transparency as a principle for regulating AI has been frequently linked to notions of explainability, intelligibility, and auditability, but it also appears in reference to the governance structures adopted in developing and implementing AI systems. Therefore, the term "transparency" has been used in both a specific sense (associated with explainability) and a wider sense. Both perspectives need to be addressed in order to understand the construction of this dimension of the regulatory framework.

4 As this article is being prepared Bill No. 2338 of 2023, which started in the Federal Senate, is at an advanced stage of debate in the National Congress, but the final version has still not been presented. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

THE MECHANISMS AND LIMITS FOR AI EXPLANATION IN HEALTHCARE

Explainability, which considers both inherently interpretable AI systems and post hoc explanations in opaque systems, has been recognized as an essential aspect of transparency in AI in general, and in AI as applied to healthcare. Healthcare is a sector in which the search for explainability is considered particularly necessary, especially for enabling the use of AI systems in clinical care activities (Herzog, 2022; Holzinger et al., 2017). Explanations of an AI system can be sought to justify decisions, improve control, improve models, or acquire new knowledge. In all these situations, the objective of the user (whether healthcare professional, patient, or regulator) is very relevant to explainability, so designing systems to provide explanations is very complex, especially when it comes to obtaining post hoc explainability (Roscher et al., 2020). The scientific and practical field of explainable AI (eXplainable Artificial Intelligence [XAI]) is expanding, and companies, standards bodies, non-profit organizations, and public institutions are currently undertaking a lot of research with the aim of creating AI systems that can explain their forecasts (Gunning et al., 2019).

Generally speaking, it is neither feasible nor necessary for one explanation to provide the entire decision-making process of a machine learning model. The explainability of an AI system is essential in situations in which there is some flaw, with regard to which a specific instance of the system needs to be determined. This is especially important when algorithmic results are used to make recommendations or decisions that would normally be subject to human discretion. To do this an explanation needs to be able to address at least one of the following points (OECD, 2019): (a) the main decision factors – indicating the important factors of a prediction made by AI, preferably ranked in order of significance; (b) the factors that determine the decisions – clarifying factors that decisively affect the result; and (c) the divergent results – clarifying why two cases that appear similar may present different results. Explanations, therefore, need to provide information about the factors that AI models use in arriving at a result, and the relative weight of each factor that can be interpreted by humans. They must also be able to provide answers to

counterfactual questions in order to know whether a factor considered in an algorithmic decision was decisive for a specific result (Doshi-Velez et al., 2019).

The most direct and accessible way to obtain explanations is to use only those algorithms that create interpretable models, such as linear regression, logistic regression, and decision trees: using interpretable AI, in other words.[5] Generally speaking, machine learning models that are developed with algorithms like these (and others that are also based on statistical approaches) can be directly interpreted by way of techniques and calculations that humans can understand. The field of AI in healthcare, however, is being increasingly driven by the use of models that have been developed by more complex machine-learning algorithms and neural networks (deep learning), which work like black boxes. For these opaque systems, it is only possible to obtain explanations using post hoc techniques – i.e., employing explainable AI. Among the various techniques currently being used in this field, the most common are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanation (SHAP), which were developed with the objective of explaining the individual predictions of models.[6] Both use an interpretable surrogate model that explains a complex (opaque) model according to the explainable AI paradigm. There are also specific methods for explaining neural network models, such as the saliency map technique (also called "pixel attribution"), which highlights relevant pixels for a given image classification, and dozens of other techniques for different types of algorithms (Molnar, 2022).

---

5 Linear regression is an algorithm that solves regression problems by defining a function that makes the intended prediction a weighted sum of the input features. Logistic regression is an extension of the linear regression algorithm for solving classification problems using a function that limits the outputs to results between 0 and 1 (probabilities). Decision trees are algorithms that split the data multiple times according to certain cutoff values in the features (inputs), creating different subsets (nodes) of the dataset (each instance in a subset), and predicting results from the result of each node. Decision trees can be used for both regression and classification problems (Molnar, 2022).

6 The LIME technique creates a local surrogate model and generates a new set of data that seeks to approximate the predictions corresponding to those of the black-box model (Ribeiro et al., 2016). The SHAP technique has the same aim. It seeks to achieve it by calculating the contribution of each resource to the prediction, based on the so-called "Shapley values": a coalition game theory method that calculates the average contribution of each member of all possible coalitions (Lundberg & Lee, 2017).

Initially considered to be almost a consensus, algorithmic explainability has been the subject of increasing controversy in the 2020s, due to recognition of the limitations of explainable AI. First, there is a trade-off between explainability and the performance of machine learning models. For an AI system to be explainable the solution variables need, by definition, to be reduced to a set that is small enough for it to be accessible to human understanding. This tends to make the use of some systems in complex problems unfeasible, which is why the prospect of demanding a detailed explanation may be incompatible with the use of AI systems that seek high predictive accuracy (London, 2019).

The explanation of AI systems is also limited by the real possibilities offered by existing mechanisms. While currently available techniques for explainability are able to provide broad descriptions of how an AI system works in a general sense, they are very superficial or unreliable for individual decisions. Another aspect to be considered is the observation that users tend to rely excessively on explanations given by explainable AI tools, often without realizing that they are not guaranteed to perform, a particularly worrying point in the healthcare area given the possible adoption of increasing numbers of AI solutions in clinical environments. As explanations are only approximations of the decision process of the opaque model, an additional source of error is created, since both the original model and the explanatory model may be wrong (Ghassemi et al., 2021).

REGULATORY PERSPECTIVES FOR TRANSPARENCY AND EXPLAINABILITY

In the scenario above, some elements might be considered for developing a regulatory framework for AI in healthcare that effectively incorporates transparency mechanisms. The proposed approach considers the current limits of the explainability of AI systems in healthcare and reinforces the conception of transparency in a broad sense with regard to the development and implementation of machine learning tools.

The trade-off between explainability and performance does not necessarily exist in any of the situations in which machine learning models are employed in practice. In AI

systems that are developed to solve structured data problems in which meaningful features are well represented (as in the context of healthcare), minimal performance differences have often been observed between complex algorithms, such as neural networks, and simpler, interpretable AI classifiers after the data have been adequately pre-processed (Rudin, 2019). It can be argued, therefore, that the use of interpretable algorithms should always be prioritized, especially when developing models for use in situations in which explanation is essential, as can happen in sensitive or critical conditions in the healthcare area.

One possible approach to regulating AI in healthcare is to require developers of black-box models to report on the performance of tested and validated interpretable models that are intended for the same uses. This would allow a direct assessment as to whether there is a trade-off between explainability and performance, which would probably encourage the use of interpretable AI algorithms whenever possible. A stronger proposal would be not to allow the use of opaque systems in high-risk situations if an interpretable system with the same level of performance exists.

From an explainability perspective, the current lack of transparency in AI in healthcare is likely to persist and even intensify over time. The field has evolved towards increasingly complex and, consequently, more opaque systems, such as large language models and foundation models. (Moor et al., 2023). This considered the path to the efficient regulation of AI in health must start by admitting that opaque systems are going to be used, instead of presupposing that there is going to be a search for humanly understandable explanations for individual algorithmic decisions. The central point must be defining the rules and conditions for using black-box systems that are in line with ethical precepts and human rights. Everything points to the fact that there will be situations in which the results of opaque models will be sufficiently positive to justify the decisions that are taken and that are based on them. If there is a robust regulatory framework for ensuring the safety and effectiveness of these systems, it is quite possible that they will be used with confidence despite any limitations as to their explainability. Opacity is, to a certain extent, a common feature

in clinical activity: Medicine traditionally adopts practices that involve mechanisms that are not fully understood but that continue to be widely used due to their proven effects, such as the use of many medications. Something similar could, therefore, happen with AI in healthcare.

However, it needs to be reiterated that the fundamental pillar for guaranteeing the transparency of AI in healthcare is rigorous and complete documentation. Meticulously registering clear and detailed information about the methods, resources, and decisions throughout the entire life cycle of AI systems, in addition to being fundamental for ensuring the reliability of these tools, must also be an essential regulatory requirement. Regulatory bodies need to have access to adequate documentation covering everything from the design, development, training, and validation of the models to the implementation and the post-implementation period (WHO, 2023). Some existing proposals can serve as a reference for defining what information should be provided and how this information needs to be organized.[7] It is necessary to require that AI systems include information about the populations used in the training data (data sources and selection of the cutoff point) and the demographics of that data, in order to allow comparison with the population in which the models will be implemented. Requirements should also include detailed information about the architecture and development of the models, in order to facilitate interpretation of the intended use in comparison with similar AI systems. Such measures can give regulators and users a better understanding of how AI systems work, considering transparency as an approach for identifying best machine learning practices (Hernandez-Boussard et al., 2020).

7 One of the most relevant initiatives currently is the Transparent Reporting of a multivariable Prediction model of individual prognosis Or Diagnosis for AI (TRIPOD-AI), an AI-specific extension of the Transparent Reporting of a multivariable Prediction model of individual prognosis Or Diagnosis (TRIPOD) statement, a standard reporting protocol that includes a 22-item checklist that is designed to improve the quality of the reporting of studies that develop, validate or update predictive models for clinical (diagnostic or prognostic) purposes. It is associated with PROBAST-AI, which was developed as an extension of the Prediction model Risk of Bias Assessment Tool [PROBAST]), a tool that assesses the risk of bias and the applicability of predictive models based on 20 questions (in four domains: participants, predictors, outcome and analysis). These tools have been designed to guide researchers and reviewers to assess the quality of studies and interpret scientifically relevant findings. (Collins et al., 2015, 2021, 2024; Wolff et al., 2019). They can also be very useful for helping regulators.

FINAL CONSIDERATIONS

AI has become one of the main technologies used in healthcare, so establishing a regulatory framework for AI in this context has assumed a central importance in contemporary society. Among the foundations for regulating AI in healthcare arising from ethics in AI, transparency is recognized as an essential dimension.

A broad understanding of the concept of transparency implies the need to make accessible all the information that forms the basis of the decisions that originate from machine learning models. There must be clarity regarding the technical and institutional contexts in which AI systems are designed, implemented, and managed, which must be expressed by way of standardized documentation. A central mechanism in the regulatory strategy, therefore, must be defining the criteria for standardizing the detailed and complete documentation of all stages in the development and implementation of AI systems considering existing lists as references.

Transparency also concerns the interpretability and explainability of AI systems. Identifying the explanation mechanisms and limits of AI in healthcare is crucial for defining the extent of the right to explanation and the regulatory requirements that must be established in this dimension. Regulatory frameworks must differentiate interpretable AI systems from explainable AI systems, recognizing that current explanation mechanisms, although capable of providing factors that may determine the decisions taken and clarify divergent results in interpretable models (interpretable AI), represent only approximations in opaque models (explainable AI). Regulatory explainability requirements, therefore, may require explanations about specific decisions in interpretable AI models but should be restricted to clarifications regarding general functioning in explainable AI systems. The requirement to compare performance with interpretable models should be considered as a regulatory requirement for systems that are based on opaque AI models. In this sense, it is necessary to learn how to deal with opacity by defining rules for the use of black-box AI systems that have shown themselves to be beneficial, in order to ensure that they are duly proven to be safe and effective.

REFERENCES

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Aith, F., & Dallari, A. B. (Orgs.). (2022). *LGPD na saúde digital*. Thomson Reuters Brasil.

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., & Madai, V. I. (2022). To explain or not to explain? — Artificial Intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), 1-18. https://doi.org/10.1371/journal.pdig.0000016

Association for Computing Machinery US Public Policy Council. (2017). *Statement on algorithmic transparency and accountability*. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, *373*(6552), 284-286. https://doi.org/10.1126/science.abg1834

Bates, D. W. (2023). How to regulate evolving AI health algorithms. *Nature Medicine*, *29*(26). https://doi.org/10.1038/s41591-022-02165-8

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of biomedical ethics*. Oxford University Press.

Bill no. 2338/2023. (2023). Provides for the use of Artificial Intelligence. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

Brownstein, J. S., Rader, B., Astley, C. M., & Tian, H. (2023). Advances in Artificial Intelligence for infectious-disease surveillance. *New England Journal of Medicine*, *388*(17), 1597-1607. https://doi.org/10.1056/NEJMra2119215

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1-12. https://journals.sagepub.com/doi/10.1177/2053951715622512

Bygrave, L. A. (2020). Automated individual decision-making, including profiling. In Christopher, K., Bygrave, L. A., Docksey, C., & Drechsler, L. (Eds.), *The EU General Data Protection Regulation (GDPR): A commentary* (Article 22, pp. 522-542). Oxford University Press. https://doi.org/10.1093/oso/9780198826491.003.0055

Collins, G. S., Dhiman, P., Navarro, C. L. A., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Calster, B. V., van Smeden, M., Riley, R. D., & Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on Artificial Intelligence. *BMJ Open, 11*(7). https://doi.org/10.1136/bmjopen-2020-048008

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., Hoffman, M. M., Kengne, A. P., Lam, E., Lee, N., Loder, E. W., Maier-Hein, L., Mateen, B. A., McMradden, M. D., Oakden-Rayner, L., Ordish, J., Parnell, R., Rose, S., Singh, K., Wynants, L., & Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*. https://doi.org/10.1136/bmj-2023-078378

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ, 350*. https://doi.org/10.1136/bmj.g7594

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI under the law: The role of explanation. *arXiv*. http://arxiv.org/abs/1711.01134

Dourado, D. A., & Aith, F. M. A. (2022). A regulação da Inteligência Artificial na saúde no Brasil começa com a Lei Geral de Proteção de Dados Pessoais. *Revista de Saúde Pública, 56*(80). https://doi.org/10.11606/s1518-8787.2022056004461

Dubber, M. D., Pasquale, F., & Das, S. (Orgs.). (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.

European Commission. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5

G7 Hiroshima Summit. (2023). *G7 Hiroshima Leaders' Communiqué*. https://www.mofa.go.jp/policy/economy/summit/hiroshima23/documents/pdf/Leaders_Communique_01_en.pdf?v20231006

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. This law addresses the processing of personal data, including in digital media, by natural persons or legal entities, whether public or private, with the aim of protecting the fundamental rights of freedom and privacy, and the free development of the personality of the natural person. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

General Data Protection Regulation (GDPR). (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable Artificial Intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gottlieb, S., & Silvis, L. (2023). Regulators face novel challenges as Artificial Intelligence tools enter medical practice. *JAMA Health Forum*, *4*(6). https://doi.org/10.1001/jamahealthforum.2023.2300

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G-Z. (2019). XAI – Explainable Artificial Intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., & Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for Artificial Intelligence in health care. *Journal of the American Medical Informatics Association*, *27*(12), 2011-2015. https://doi.org/10.1093/jamia/ocaa088

Herzog, C. (2022). On the ethical and epistemological utility of explicable AI in medicine. *Philosophy & Technology*, *35*(50). https://doi.org/10.1007/s13347-022-00546-y

Hinton, G. E. (2018). Deep learning-A technology with the potential to transform health care. *JAMA*, *320*(11), 1101-1102. https://doi.org/10.1001/jama.2018.11100

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv*. https://doi.org/10.48550/arXiv.1712.09923

International Bioethics Committee. (2017). Report of the IBC on Big Data and health. *United Nations Educational, Cultural and Scientific Organization*. https://unesdoc.unesco.org/ark:/48223/pf0000248724

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1,* 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal, 34*(1), 189-218. https://doi.org/10.15779/Z38TD9N83H

Khorsandi, S. E., Hardgrave, H. J., Osborn, T., Klutts, G., Nigh, J., Spencer-Cole, R. T., Kakos, C. D., Anastasiou, I., Mavros, M. N., & Giorgakis, E. (2021). Artificial Intelligence in liver transplantation. *Transplantation Proceedings, 53*(10), 2939-2944. https://doi.org/10.1016/j.transproceed.2021.09.045

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57. https://dl.acm.org/doi/10.1145/3236386.3241340

London, A. J. (2019). Artificial Intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15-21. https://doi.org/10.1002/hast.973

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the International Conference on Neural Information Processing Systems* 31, 4768-4777. https://dl.acm.org/doi/10.5555/3295222.3295230

Matheny, M. E., Israni, S. T., Ahmed, M., & Whicher, D. (Orgs.). (2022). *Artificial Intelligence in health care: The hope, the hype, the promise, the peril*. National Academy of Medicine.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. (2022). Interpretability and transparency in Artificial Intelligence. In C. Véliz (Org.), *The Oxford handbook of digital ethics* (pp. 378-410). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.20

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2ª ed.). https://christophm.github.io/interpretable-ml-book/

Monteiro, R., Machado, C. V., & Silva, L. (2021). The right to explanation in Brazilian Data Protection Law. *International Journal of Digital and Data Law, 7*(1), 119-136. https://ojs.imodev.org/?journal=RIDDN&page=article&op=view&path%5B%5D=406

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical Artificial Intelligence. *Nature, 616*. https://doi.org/10.1038/s41586-023-05881-4

Nature, E. (2023). AI's potential to accelerate drug discovery needs a reality check. *Nature, 622*, 217. https://doi.org/10.1038/d41586-023-03172-6

Organisation for Economic Co-operation and Development. (2019). *Artificial Intelligence in society*. https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Nova York, NY, Estados Unidos da América, 22, 1135-1144. https://dl.acm.org/doi/10.1145/2939672.2939778

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access, 8*, 42200-42216. https://doi.org/10.1109/ACCESS.2020.2976199

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5). https://doi.org/10.1038/s42256-019-0048-x

Sahni, N., & Carrus, B. (2023). Artificial Intelligence in U.S. health care delivery. *New England Journal of Medicine, 389*(4), 348-358. https://doi.org/10.1056/NEJMra2204673

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233-242. https://doi.org/10.1093/idpl/ipx022

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *Npj Digital Medicine, 3*(1). https://doi.org/10.1038/s41746-020-0221-y

Topol, E. J. (2019). High-performance medicine: The convergence of human and Artificial Intelligence. *Nature Medicine, 25*(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: Beyond the black box. *BMJ, 364*, l886. https://doi.org/10.1136/bmj.l886

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51-58. https://doi.org/10.7326/M18-1376

World Health Organization. (2021). *Ethics and governance of Artificial Intelligence for health: WHO guidance*. https://apps.who.int/iris/handle/10665/341996

World Health Organization. (2023). *Regulatory considerations on Artificial Intelligence for health*. https://iris.who.int/handle/10665/373421

World Health Organization. (2024). *Ethics and governance of Artificial Intelligence for health: Guidance on large multi-modal models*. https://iris.who.int/handle/10665/375579

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial Intelligence in healthcare. *Nature Biomedical Engineering, 2*, 719-731. https://doi.org/10.1038/s41551-018-0305-z

Part 2

QUALITATIVE RESEARCH

Methodological notes

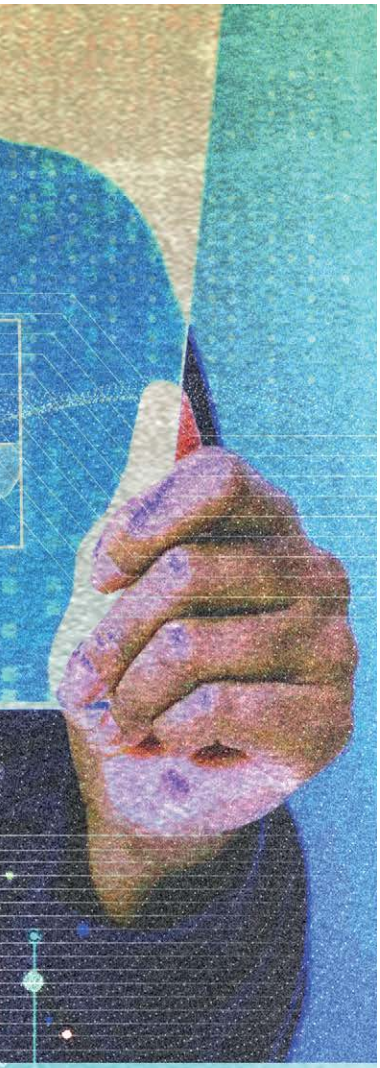
Graziela Castello,[1] Monise Picanço,[2] Priscila Vieira,[3] and Rodrigo Brandão[4]

1 Social scientist and coordinator of Qualitative Methods and Sectoral Studies in the Regional Center for Studies on the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center (NIC.br).

2 Ph.D. and master's degree in Sociology from the University of São Paulo (USP). Associate researcher and project coordinator at the Brazilian Center of Analysis and Planning (CEBRAP). She also teaches methodology and research design at the Paulista School of Informatics and Administration (FIAP) and cebrap.lab.

3 Sociologist with a master's and Ph.D. degree in Sociology from USP. Researcher and project coordinator at the Development Center of CEBRAP.

4 Ph.D. candidate in Sociology at USP, holds a master's degree in Political Science from USP, and a bachelor's degree in Social Sciences from USP. Qualitative Methods and Sectoral Studies researcher at Cetic.br|NIC.br.



This chapter presents the methodological procedures adopted in the study "Artificial Intelligence in health: A qualitative diagnosis of the scenario in Brazil". The analysis of its results can be found in the next chapter. The study is an exploratory qualitative investigation involving key stakeholders in the health sector in Brazil, who are at the forefront of the knowledge on the subject in their different fields of activity. The investigation aimed to map out the current debates on the public agenda, collect perceptions regarding the opportunities, challenges, and risks for developing and adopting tools and solutions based on Artificial Intelligence (AI) in healthcare, and identify current initiatives and practices in the Brazilian context. This chapter also presents the study's general objectives, research design, data collection instruments, and methodology for analyzing the results.

THE STUDY'S GENERAL OBJECTIVES AND THE PLANNED DESIGN

Using complex health data and multiple sources — electronic health records, imaging studies, and genomic and physiological data, for example — AI has great potential for facing up to the growing challenges in the health sector, such as the continuous increase in costs, the lack of specialists, ongoing epidemiological and demographic changes, population aging, and others. In this sense, the use of AI-based tools represents a promising development for leveraging and accelerating the production of scientific knowledge and for developing and implementing public health policies on a large scale (Brazilian Academy of Sciences [ABC], 2023).

Given this scenario, the study's objective was to map the current stage of AI development in the healthcare sector in Brazil (where we are and how we are doing) based on an exploratory diagnosis whose specific objective was to gather information on: (a) the opportunities for the country with the advent of this technology in the healthcare area; (b) the challenges posed for taking advantage of these opportunities; (c) the potential risks to be managed; (d) the ongoing research agendas, policies, and initiatives in the country; and (e) the perceived impacts of the use of AI tools in clinical practice.

Due to its exploratory nature and the objective of problematizing the potential, risks, and perspectives for Brazil with the advent of AI in healthcare, the study (which had a qualitative nature) was based on in-depth interviews (IDI) with strategic actors in the country that are in the intersection of AI and health. For this study's purposes, they were considered privileged informants and opinion makers on the topic (stakeholders).

In qualitative research, "informant" refers to a person selected to provide detailed information and insights on a topic. In the study, "informants" were asked to provide information on AI in healthcare, and they will be referred to as "interviewees" in the analysis of the collected data. The selection of privileged informants (who are at the forefront of the current debate on the topic) was based on the premise of exploring existing experiences in the country to substantively investigate the opportunities, challenges, and barriers based on a concrete situation – the current stage of development of AI in the healthcare sector in Brazil. This qualitative study did not gather information from informants in the healthcare sector from locations and institutions without AI-related activities. This qualitative study, therefore, is a picture of the Brazilian scenario of those directly connected to AI in healthcare, and, in this sense, it has the intentional bias of investigating the opportunities and challenges based on the perceptions and opinions of individuals closely involved with this agenda. As a methodological choice, this approach enables a deeper analysis of the topics investigated based on the situations and experiences of stakeholders affected by the topic. By adopting such a strategy, it is possible not only to map opportunities and challenges but also to qualify the different levels of priority among them for the timely development of AI in the country's healthcare sector, considering the expertise of those involved in the topic.

Based on the proposed design, the criteria for selecting interviewees became a central part of the study and took into account the following characteristics: (a) diversification in the segments they represent in the health area; (b) holding a leadership and/or management position in the organization in which they work, or being a specialist; (c) significant

involvement in the public debate, considering scientific production, participation in public events on the topic, and exposure in the media related to the theme.

Considering these criteria, a necessary step was mapping potential interviewees based on a review of publicly available content, scientific production, presence at events, exposure in the media on the topic, an active search in forums and institutional websites, and indications from publicly recognized actors. The mapping out process resulted in an initial list of 93 names of potential interviewees, drawn from various segments in the healthcare area that use AI. The second stage in the selection process was considering the diversity of segments represented in the study, with the allocation of a minimum number of five interviews per segment.

Five segments were defined for selecting interviewees based on the nature of the main activities of the organizations in the healthcare area in which the potential interviewees worked, such as research, clinical practice, the development of solutions, and the promotion of public policies, all of which converge with the definition found in the Digital Health Strategy for Brazil 2020-2028 (DHS) on digital healthcare actors (Ministry of Health [MS], 2020).[5] The selection of respondents in each segment also prioritized the diversity of the organizations; therefore, it was established that only one interview per organization would be conducted. From the initial list of 93 names of potential interviewees, 28 specialists were selected according to the criteria that had been defined. They were distributed as shown in Table 1.

It is essential to highlight that only the frontline healthcare professionals were selected based on recommendations by the interviewed specialists (stakeholders) because AI tools are still not widely used in the Brazilian healthcare sector, and there is not enough public information on the subject. In this sense, this subsample was constructed for convenience's sake.

5 It is important do highlight that "patient associations" and "citizens," two groups found in digital healthcare, as defined by the DHS, were not the subject of investigation in this study.

TABLE 1 – DESCRIPTION OF THE SEGMENTS INVESTIGATED AND THE TOTAL NUMBER OF INTERVIEWS

Segments	Segment description	Equivalence in the DHS[6] (actors in digital healthcare)	Total number of interviews
Academia	Universities, research centers, think tanks, reference centers	Technical-scientific associations, universities, and training centers	6
Healthcare facilities (public and private)	Hospitals, clinics, laboratories	Service providers for the health system	6
Government	Ministries, government departments and agencies	Ministry of Health; agencies; state and municipal health departments	5
Market	Technology companies, startups, health plans	Industry and technology sector; health system sources of payment	6
Healthcare professionals (at the frontline)	Doctors and nurses who use AI tools at the frontline	–	5
Total			**28**

SOURCE: PREPARED BY THE AUTHORS.

The fieldwork was carried out between August and December 2023. The interviews lasted an average of one hour, were conducted remotely with the recruited interviewees through previously scheduled video calls and were recorded and transcribed for later coding of the material.

All recruited participants were informed about the nature and objectives of the study and how the collected data would be treated and used. Before the interviews, the informants signed a Consent Form detailing this information. The research followed the guidelines of the Brazilian General Data Protection Law (LGPD) (Law No. 13.709/2018) and research ethics protocols, which would allow participants to withdraw at any time and would ensure anonymity and information confidentiality. Participation was voluntary, with no payment or any other type of material incentive provided for the interviewees' participation.

6 For further information about the actors in digital healthcare as defined by the DHS, see MS (2020).

COLLECTION INSTRUMENTS

The same semi-structured script, divided into two modules, was used to conduct the interviews with stakeholders from the academia, market, government, and healthcare facilities groups. The first module prompted questions about the current AI scenario as applied to Brazil's healthcare. It sought to collect perceptions of the current stage and the opportunities, challenges, risks, and priorities. Additionally, it aimed to gather information about ongoing AI initiatives in healthcare in the institutions to which the interviewees were linked to understand the nature of the practices, their stage of development, their potential, and the difficulties faced in their design and implementation. The second module was based on the seven DHS priorities, which are essential to the promotion of digital health in Brazil (MS, 2020). Its questions were only applied when the topic did not appear spontaneously during the execution of the first module of the script (Appendix 1).

The informants in the "frontline healthcare professionals" group were interviewed using another script, which was also semi-structured but with substantial differences, and whose focus was on understanding the inclusion of AI tools in daily work routines. The script, therefore, prompted questions about the application of AI in clinical practice, the adoption process, training, trust, interpretability, the relationship with patients, challenges, risks, and prospects. Since the research with this group had a different analytical objective and was conducted using a different collection instrument, the data from these interviews were analyzed separately. The complete script used with frontline healthcare professionals can be found in Appendix 2 of this chapter.

METHODOLOGY USED FOR ANALYZING THE RESULTS

After conducting the interviews and transcribing them, the analytical phase was carried out using content analysis methodology to systematize the data from the IDI through the coding process. This process systematically categorizes qualitative data to identify relevant patterns and themes in the interview transcripts.

Content analysis is a qualitative methodology that seeks to develop a systematic and objective description of manifest

messages,[7] that is, regarding the presented content. Its purpose is to critically analyze the materials presented without losing sight of the need to develop an analysis with valid inferences and replicable procedures. This method originated in the 1970s and uses an analysis logic that is often deductive (it verifies and proves by testing hypotheses) or abductive (it investigates causes and/or mechanisms to explain the occurrence of a particular phenomenon) (Krippendorff, 1980; Neuendorf, 2002).

When using coding as a systematization and analysis procedure, this methodology employs the following steps in its operationalization: Pre-analysis of the materials, development of a coding plan, coding of a sample of the data for validation purposes, a review of the codes, coding of the material, and an analytical description of the main results (Bardin, 1977). It is particularly important for this analysis to establish a robust coding plan a priori, which is established before all the material is systematically analyzed.

For the pre-analysis of the materials in this study, a cross-sectional reading of all the material was conducted, with an initial systematization of the main highlights. Subsequently, a systematic reading of a sample of interviews was performed, selecting one interview from each of the segments defined in the study. This step was crucial for developing the coding plan, involving the organization and categorization of the information collected in the interviews, built from a collective discussion considering the research objectives.

The coding plan that was developed has two layers of systematized information. The first is based on the interview script and its questions to locate the prompts used in the interviews. The second layer of information is formed of analytical codes, that is, themes and sub-themes identified in the pre-analysis of the materials.

After establishing the preliminary coding plan, a test was carried out based on the coding of a small number of interviews. The revision of the coding plan after the test incorporated two additional codes, which validated and demonstrated the robustness of the initial plan applied to the remaining material.

7 This methodology differs from discourse analysis, for example, which often focuses on the meanings of the statement and what it reveals about its context (Orlandi, 1999).

The codification was carried out using Atlas.ti software,[8] which enabled the qualitative data to be organized and analyzed, as well as the segmentation of interview texts into specific excerpts, facilitating the assignment of codes to these segments according to the identified dimensions and themes. Furthermore, Atlas.ti offers several features that improve the coding and analysis process, such as creating and managing codes, visualizing relationships between different themes, and generating reports summarizing the coded information. Therefore, using this software provided a robust systematization of the collected data, which meant that the results could be analyzed in an organized and consistent manner.

As a result, a detailed database was created, identifying the interview excerpts corresponding to each of the identified dimensions and their profile characteristics. This systematization of the material by coding themes supports the analysis of the results presented in the next chapter of this publication.

8 Find out more: https://atlasti.com/

REFERENCES

Brazilian Academy of Sciences. (2023). *Recomendações para o avanço da Inteligência Artificial no Brasil: GT-IA da Academia Brasileira de Ciências*. https://www.abc.org.br/wp-content/uploads/2023/11/recomendacoes-para-o-avanco-da-inteligencia-artificial-no-brasil-abc-novembro-2023-GT-IA.pdf

Brazilian Ministry of Health. (2020). *Estratégia de Saúde Digital para o Brasil 2020-2028*. https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf

Bardin, L. (1977). *Análise de conteúdo*. Edições 70.

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. This law addresses the processing of personal data, including in digital media, by natural persons or legal entities, whether public or private, with the aim of protecting the fundamental rights of freedom and privacy, and the free development of the personality of the natural person. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology* (2nd ed.). SAGE.

Neuendorf, K. A. (2002). Defining content analysis. In *The Content Analysis Guidebook* (pp. 1-26). SAGE.

Orlandi, E. P. (1999). *Análise de Discurso: princípios e procedimentos*. Pontes.

APPENDIX 1 – SCRIPT ON AI IN HEALTHCARE IN BRAZIL APPLIED TO THE STAKEHOLDERS (ACADEMIA, HEALTHCARE FACILITIES, PUBLIC AUTHORITIES, AND THE MARKET)

MODULE 1.1: BRAZILIAN SCENARIO

To begin talking about AI in healthcare, let us consider the scenario in Brazil.

1. In your opinion, what is the current stage of Artificial Intelligence applied to healthcare in Brazil?
2. In your opinion, what opportunities does Artificial Intelligence create to optimize Brazil's healthcare sector? (EXPLORE: When I say to optimize the healthcare sector, I refer to opportunities for improving services, increasing care scale, and reducing sector costs)
3. And how prepared do you think the country is to take advantage of the opportunities and manage the risks associated with the advance of AI in the healthcare sector? Why?
4. Also, along the same lines, what should the country's priorities be for taking advantage of the opportunities that AI can bring to the healthcare sector, in your opinion? Which areas, sectors, and segments would have the most significant potential? Where should we invest?
5. And what are the possible risks associated with the use of AI in the healthcare sector?
6. What are the main bottlenecks, challenges, and barriers to the development of AI in the healthcare sector, in your opinion? Which challenges are common to the different countries and contexts, in your opinion? What are the specific challenges in Brazil?

MODULE 1.2: ONGOING PRACTICES

Now, we are going to talk about the AI practices and uses in healthcare that are already being employed in your institution, thinking specifically about your routine and that of your institution with the application of AI in the healthcare area.

7. (RESEARCH CENTERS/ HEALTHCARE FACILITIES/ COMPANIES) What research agendas, projects,

technologies, and/or solutions in AI in healthcare is your institution developing? To what end?

a. What is the current stage of development of these projects in AI in the healthcare area?
    i. In the specific case of your institution, have you looked for or created any incentives for developing the area? How did you do it?
b. What potentialities/opportunities do these projects have/offer, in your opinion?
c. What are the main difficulties/obstacles you have faced in these fields of activity?
d. Finally, how do you deal with the risks that AI poses?

8. (PUBLIC MANAGERS) What are the main proposals, agendas, and actions you are currently debating regarding possible advances in AI in healthcare?
    a. What is the current stage of development of these projects in AI in the healthcare area?
        i. In the specific case of your institution, have you looked for or created incentives to develop the area? How did you do it?
    b. What potentialities/opportunities do these projects have/offer, in your opinion?
    c. What are the main difficulties/obstacles you have faced in these fields of activity?
    d. Finally, how do you deal with the risks that AI poses?

MODULE 2: 7 DHS PRIORITIES

Considering that the potentialities envisioned with the advent of Artificial Intelligence in the healthcare sector depend on the organization, processing, and management of a vast range of data and information, we would like to explore your perceptions of some specific topics.

M2.1 In your opinion, what are the main infrastructure challenges that Brazil will need to face to maximize the potential

opportunities of AI in the healthcare sector? Considering facilities, technologies, and resources, what should be the country's investment priorities in terms of infrastructure?

M2.2 Regarding the human resources needed to develop and take full advantage of AI applications in healthcare, do you think healthcare professionals in Brazil are ready to deal with these new technologies? What qualifications and/or skills will they need? How can professionals be prepared for this possible future?

M2.3. Taking advantage of the potential of AI involves ensuring the broad, diverse, and large-scale use of different data sources. Considering this, what strategies could be implemented to promote the interoperability of information systems? What precautions should be taken to promote this?

a. What incentives could be created to guarantee data integration and good governance?
b. How has your institution been addressing these issues, including efforts to achieve greater data interoperability with other institutions?

M2.4. Still thinking about the environment necessary for the development of AI in the healthcare area, could you tell me, in your opinion, what are the main challenges the country faces in relation to:

a. Regulation. What is needed? What is the country like regarding this?
b. Ethics. What are the main ethical questions that must be addressed for developing AI in a safe and beneficial way in the healthcare area in the country?

M2.5. How much do you know about the RNDS – National Health Data Network (*the national health interoperability platform of the Programa Conecte SUS - MS*)? What do you know about it? (*FOR THOSE WHO KNOW ABOUT IT*). As the country's data integration platform, would the RNDS have the potential to meet the demand that AI in healthcare generates? How would it do this?

M2.6. Finally, let's quickly talk about the healthcare users: The citizens.

a. What user rights should be safeguarded most when discussing AI? (ONLY ENCOURAGE

INTERVIEWEE IF NECESSARY). How should matters of confidentiality be dealt with? How can transparency and ethics be guaranteed in the processes?

b. What policies does your institution currently have for dealing with these issues? Are they sufficient? What could be improved? What works well?

APPENDIX 2 – AI GUIDELINES IN HEALTHCARE IN BRAZIL APPLIED TO FRONTLINE HEALTHCARE PROFESSIONALS

GENERAL PRESENTATION

1. Initially, you should introduce yourself and talk briefly about what you currently do (where you work, your medical specialty, and what type of clinical care you offer)?

APPLICATION OF AI IN CLINICAL PRACTICE

2. Can you tell us how you use/have used AI tools in your clinical practice? [Explore: Which tools they are, who developed them, their objectives, and how often they are used].
3. Can you tell us about/describe in detail how this tool works and how you apply it in your day-to-day work? At what specific moment or in what situation do you apply it? How do you use it to make clinical decisions? [availability of information, clinical report, diagnosis?]

PROCESS OF ADOPTING/STARTING TO USE AI IN CLINICAL PRACTICE

4. How was the process of adopting these tools in your clinical practice? Has anything changed in the way you use them today?
5. Why did you decide to adopt these tools? [Explore: Did you have any kind of incentive or imposition from the institution where you work? Was it your own initiative? Were you inspired by the practice of someone/some professional/some institution?]
6. Did you have any difficulties in the beginning? Which?
7. Did you have any training? How did you learn to use the tool?
8. Considering what your clinical practice was like before, what changed when you started using AI tools?
    a. What are the main impacts of adopting AI in your clinical practice?

PATIENT

9. Thinking about your relationship with your patients, did anything change in your relationship with them after adopting these AI tools?
10. Is the patient informed about the use of this tool in his/her care/treatment? What information does he/she receive, and at what point during the care/treatment? If they do not receive this information, why is he/she not informed? [Explore: What are the guidelines of the organization where you work on this issue?]
11. Can patients refuse to accept the tool used in their treatment/care? How does this work? [Explore: What are the guidelines of the organization you work for on this issue?]
12. What type of patient data is recorded? Does the patient know how this data is used? [Explore: What are the guidelines of the organization you work for on this issue?]
13. Where is this data recorded, and how is it used? Is this data shared? If so, how?

THE TOOL: POTENTIAL, IMPROVEMENTS, CHALLENGES, AND RISKS

14. Thinking about your clinical practice, what are the advantages of this tool? What kind of problems does it help solve/minimize?
15. And from the patient's perspective, what is this tool's advantage for promoting their health and well-being? In your opinion, is there any harm to the patient from adopting this tool?
16. What are the determining factors for the good use of this type of tool, in your opinion? And what are the main barriers, the current obstacles, to it being used well?
17. Has this tool ever caused any difficulties/problems for you? Which? And in what way?
18. What enhancements/improvements could be made to the tools that you use to make them an even more beneficial resource?
19. Do you identify any risks associated with the use of this AI tool? What risks? To whom?

20. Can you imagine strategies to eliminate or reduce these risks? Which?

TRUST:

21. Do you trust the results generated by AI technology?
22. Do you fear that it might lead you to make incorrect decisions?
23. In your opinion, who should be held responsible for possible errors or mistakes in diagnoses and treatments recommended by these technologies? Why?

INTERPRETABILITY:

24. Is the way this technology works clear to you?
25. Has any patient ever asked you how this technology works?
    a. If "yes," did you feel prepared to provide that explanation? Which strategies did you use to give your explanation?
    b. If "no," do you feel prepared to provide that explanation? How would you approach it?

FUTURE PROSPECTS

26. Finally, when you consider the development of Artificial Intelligence technologies in healthcare, what is the future for healthcare professionals, in your opinion, especially for frontline healthcare professionals?
27. What policies or actions should be prioritized in Brazil regarding ensuring that healthcare professionals at the frontline use AI technologies appropriately?

Artificial Intelligence in healthcare: A qualitative diagnosis of the Brazilian scenario

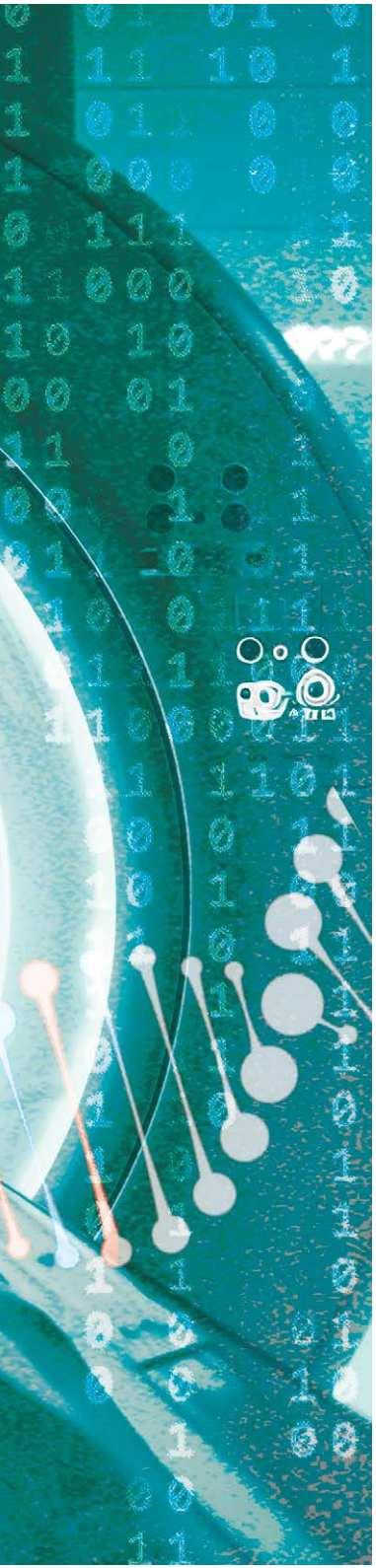
Graziela Castello,[1] *Monise Picanço,*[2] *Priscila Vieira,*[3] *and Rodrigo Brandão*[4]

1 Social scientist and the coordinator of Qualitative Methods and Sectoral Studies at the Regional Center on the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center (NIC.br).

2 Ph.D., with a master's in Sociology from the University of São Paulo (USP). Associate researcher and project coordinator at the Development Center of the Brazilian Center for Analysis and Planning (CEBRAP). Professor of Research Methodology and Design at the Paulista School of Informatics and Administration (FIAP) and Cebrap.lab.

3 Sociologist with a master's and a doctorate in Sociology from USP. Researcher and project coordinator at the Development Center of CEBRAP.

4 Ph.D. student in Sociology at USP, with a master's in Political Science from USP, and a graduate degree in Social Sciences from USP. He is a researcher with Coordination of Qualitative Methods and Sector Studies at Cetic.br|NIC.br.



INTRODUCTION

Emerging digital technologies and the possibilities arising from process automation have the potential to revolutionize and bring significant benefits to the health sector at various stages, such as: Prevention, diagnosis, prognosis, care, the treatment of individuals, and the management of healthcare systems and facilities, thus optimizing the work of professionals, resources and established operations and routines. In this context, Artificial Intelligence (AI), using complex health data from multiple sources — such as electronic medical records, imaging studies, and genomic and physiological data — has great potential to address the growing challenges of digital transformation in the healthcare sector. These challenges include constantly rising costs, a shortage of professionals, ongoing epidemiological and demographic changes (such as an aging population), and others (Brazilian Academy of Sciences [ABC], 2023, p. 7-8).

In this sense, the use of AI tools represents a promising strategy for developing and implementing large-scale public health policies, as well as boosting and accelerating scientific knowledge production. In addition to the potential benefits, the development and adoption of AI technologies also pose risks that need to be mitigated, especially in the healthcare sector.[5]

Given this scenario, it is crucial to understand the "state-of-the-art" regarding the adoption of AI in Brazil's healthcare sector: Looking at the opportunities and challenges that need to be addressed in the country, and at what stage of development and use these tools are in healthcare establishments and among professionals working in the sector. This includes questions on: (a) the progress being made by scientific production in Brazil on the subject, and the central debates and agendas; (b) the limits and possibilities offered by the adoption of AI tools for managing and improving healthcare systems (public and private); and (c), fundamentally important, the adoption motivators, the clinical significance, and the implications for the professionals who use these tools.

5 For examples of discussions on the possible risks of AI in healthcare, see Challen et al. (2019) and Adler-Milstein et al. (2022).

This chapter presents the results of the study "Artificial Intelligence in healthcare: A qualitative diagnosis of the Brazilian scenario" carried out by Cetic.br, a department of NIC.br, which is associated with the Brazilian Internet Steering Committee (CGI.br), in partnership with CEBRAP. This exploratory study aims to map ongoing experiences, gather the perceptions of strategic players, and collect information on the current stage of AI development in the Brazilian healthcare sector. It investigates, in particular, the opportunities, risks, and challenges involved in harnessing the potential of these technologies in this sector.

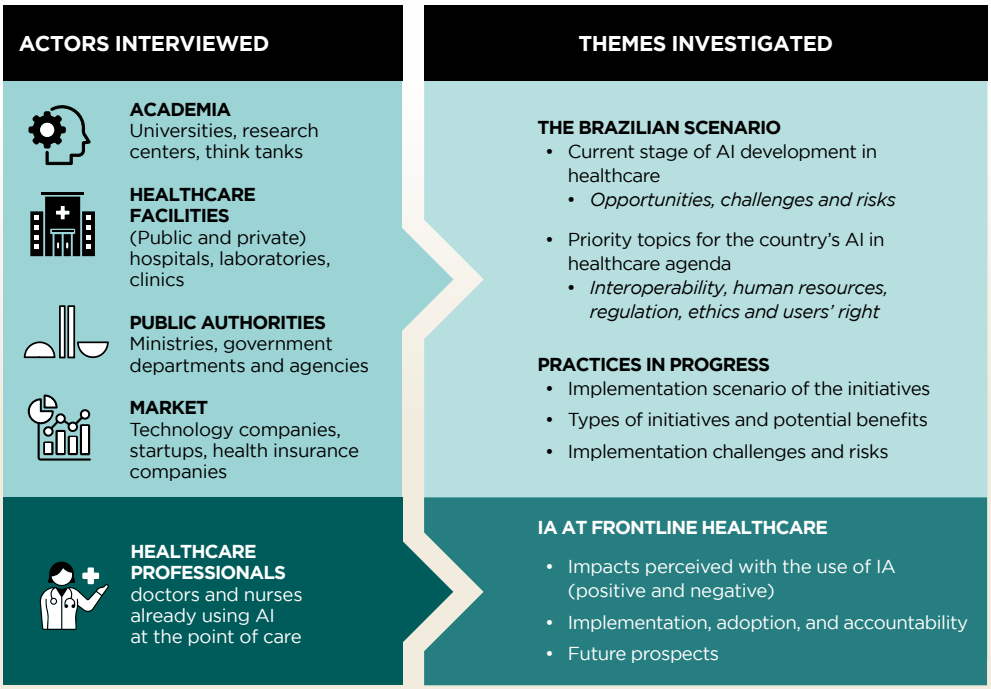
The research consisted of a qualitative survey with various stakeholder profiles involved in developing and implementing AI in Brazil's healthcare sector. A total of 28 in-depth interviews were conducted, each lasting approximately one hour, with managers and experts in the subject from academia and research centers, the public sector and the market, and healthcare establishments (public and private), as well as with healthcare professionals who use these technologies to care for patients at the frontline of healthcare. The Chapter "Methodological Notes" in this publication provides details of: (a) the study design; (b) the methodological definitions; (c) the respondent profiles, the selection criteria, and the total number of interviews carried out per profile; (d) the data collection instruments; and (e) the techniques used for coding and analyzing the interviews.

Seeking a variety of perspectives on the topic, the study aimed to recognize the main issues on the current agenda of AI as applied to healthcare in Brazil and, thus, map the emerging ideas, imminent concerns, and future expectations. In this way, the research explored perceptions of the current stage of development of AI in the country's healthcare sector, as well as the opportunities, challenges, and risks that are particular to the reality of Brazil. We also sought to gather information on initiatives and practices underway in different healthcare areas and on what is understood by data management, information security, ethics, regulation, and other topics relevant to this discussion. As a result, this study has identified an intense debate about AI in the healthcare field and the fact that the ecosystem enjoys a lot of optimism

regarding the potential benefits of this technology. It also identified that there are complex views and divergences between the different stakeholders, but mainly that there are many overlaps and convergences, which will be highlighted.

This chapter describes the main results of this empirical research and has four sections, in addition to this brief presentation. The following sections present the results of the interviews with stakeholder*s*, systematized along three lines, as shown in Figure 1. The "Brazilian scenario" Section presents perceptions of the current state of AI in healthcare in the country and views on the opportunities, difficulties, and risks associated with using these tools in the Brazilian reality. The "Ongoing practices" Section describes the initiatives for formulating, developing, and implementing AI in healthcare in the country, as mapped in the study across the different groups studied: Academia, the market, the public sector, and healthcare facilities. The "AI at the frontline of healthcare" Section provides data on the use of these technologies by healthcare professionals in patient care situations (at the frontline of healthcare). The chapter ends with a "Final considerations" Section highlighting the study's main findings.

FIGURE 1 - ANALYSIS STRUCTURE: PROFILE OF THE ACTORS INTERVIEWED AND THEMES INVESTIGATED



SOURCE: PREPARED BY THE AUTHORS.

THE BRAZILIAN SCENARIO

This section presents the stakeholders' perceptions of AI as applied to healthcare in the Brazilian context. The aim is to present a multifaceted panorama of this issue and identify the ideas under debate, as well as the wishes and prospects formulated by several of the country's key players on this agenda.

It describes the understanding of the current stage of the adoption of AI in healthcare in Brazil and presents views on the opportunities, difficulties, and risks associated with the use of AI-based tools in the Brazilian reality. Each of these themes is a topic in this section, which seeks to portray the perceptions of AI in healthcare in Brazil. The final topic is an analysis of specific questions regarding the Digital Health Strategy (DHS) related to the AI debate. The main analytical highlights are systematized at the end of each topic.

Emphasis was initially placed on the most frequently vocalized consensuses, themes, and issues raised in the interviews, i.e., the perceptions that are common to all stakeholder groups, since they were the most convergent in the different groups; the particularities we observed in specific groups will, however, be highlighted as well. Finally, possible "gaps" were also explored, i.e., issues or themes that appeared with less intensity than expected, according to the relevant literature.

CURRENT STATE OF AI IN HEALTHCARE: ADVANCES AND GAPS

This section gives an overview of the current state of AI as applied to healthcare in Brazil. At the beginning of each interview, we tried to gather perceptions of the current scenario and understand the progress and gaps peculiar to Brazil's reality. We also investigated perceptions of the country's capacity and preparedness to absorb these advances and fill the gaps.

These initial questions provided a panorama of many relevant themes and issues, which were explored and detailed throughout the interviews. The following issues were identified: (a) widespread optimism regarding the possibilities of AI for health in Brazil; (b) some concern that AI is a "fad," to the point that many technologies that are not AI are being classified as such; (c) a common perception that AI is still at an early stage in the health area in Brazil; (d) the existence of just a few initiatives, which are fragmented and disjointed; (e) initiatives generally originate in the market, since the public sector have little capacity to influence the agenda due to bureaucracy, or the need for greater care in public practice; finally, (f) specific areas, despite their general incipience, are at a more advanced stage of development, such as radiology. At first, these aspects will be broadly presented and explored in depth in subsequent topics.

When asked about the current state of AI applied to healthcare in Brazil, the study's interviewees expressed their many expectations and great enthusiasm for what can be developed in the future. "I can look ahead, and it's a very promising future," said a healthcare facility stakeholder.

There is generally great optimism and a widespread perception that the use of this technology tends to offer numerous benefits to the country's public and private healthcare systems. There is also a consensus that the use of AI in healthcare has enormous potential, both in management and medical care, especially in supporting competent professionals in their decision-making. In addition to reducing costs, efficiency and the scaling up of health services are also expected from tools that apply AI in the most timely and promising way.

People are very excited. We see a lot of students who are extremely enthusiastic about AI. [...] There's a lot of interest from the population and workers [in the field].
(ACADEMIA STAKEHOLDER)

Regarding the almost generalized optimism we observed, it is worth noting that some interviewees warn that this tends to produce a trivialization or vulgarization of the term "Artificial Intelligence" in healthcare. Although solutions that use algorithms or even AI in simple automation processes are presented as something new, they do not necessarily represent innovative uses of AI based on machine learning or deep learning for supporting management or health care. One informant noted that the current enthusiasm has created a "fad": Everybody needs to claim they use AI in their professional practice, even if, in doing so, they misuse the term or use it in an overestimated way.

People insert algorithms or new technologies, and it has become common to say, "I use AI." They've vulgarized the term AI in my view. There's a big difference between using predictive algorithms and adopting algorithms that learn on their own and can actually deliver something or make decisions to improve some delivery. I see that we're still in the process of figuring out the best way to fit the concept and use of AI into this health journey.
(MARKET STAKEHOLDER)

Today, AI is assuming large proportions, so there's the question of fads. Everyone is talking about it.
(HEALTHCARE FACILITY STAKEHOLDER)

I'd say that a lot of people are engaged in adopting AI as a tool they use because there's a perception on the part of those who build things — the builders — that is, "Look, I need to fit AI in here." So, I'd say that everyone is committed to using AI, even if they don't yet know what it's for or why. That's the fear I see in the market.
(MARKET STAKEHOLDER)

I just want to add something: There's a lot of confusion about what AI is, what advanced statistics are, what a bot is, and what is merely automation. So, we need to separate them first.

(MARKET STAKEHOLDER)

The great expectations for the future concerning the development of AI-based technology solutions for the healthcare sector in the country are linked to an assessment of the current scenario. There is a consensus that the application of AI in healthcare in Brazil is still in the early stages. Terms such as "initial," "beginning," "embryonic," and "experimental" were frequently mentioned, as were metaphors such as "a baby crawling" or "learning to walk."

It's very embryonic, isn't it? [...] But even at the embryonic stage, it's already delivering a lot of interesting value. [...] I think that even though it's in the embryonic stage, a lot of great things have already been done, and a lot of things are being developed. Goodness! There's going to be a boom in the future!

(HEALTHCARE FACILITY STAKEHOLDER)

On a scale of 0 to 10, and being very pragmatic, I'd say that Brazil is going for 2 or 3: It's not even crawling yet. It's watching how things go.

(ACADEMIA STAKEHOLDER)

We saw the first wave, which was one of experimentation; a series of experiments all over the country. You see people in healthcare testing apps and some startups experimenting with apps for clinical purposes to actually help with diagnosis and patient monitoring. And there's even the back office to improve the automation of the revenue cycle and reporting tasks. I think the second wave was applying these solutions in operations using AI to improve the detection of pulmonary nodules using MRI scans, and in chest CT scans, for example. You also have apps used to reduce loss and waste in the pharmaceutical area and in procedures. So, I'd say that we're now moving from this stage to actually implementing these technologies in Brazil.

(MARKET STAKEHOLDER)

Despite the consensus that the country is at an early stage in the development and application of AI in healthcare, some interviewees pointed out that the gap between Brazil and advanced countries in this area is not so wide because the application of AI in healthcare has occurred at a slower pace than in other areas. Incorporating these tools into clinical practice and diagnostic support is facing more restrictions and needs to undergo more tests and be subject to stricter regulations. Even the countries that are most advanced in the technological development of these tools are still in the initial stages of implementation and learning how to deal with the medical and regulatory implications of using these

technologies in practice. The following quotes illustrate this comparative interpretation:

> In Brazil, as in other countries around the world, it's still at a very early stage.
>
> *(ACADEMIA STAKEHOLDER)*

> In healthcare, the same challenges we are facing in a large hospital here in São Paulo are the same as those we encounter in Houston because of the regulations. So, I believe that of all the sectors [where AI is applied], the smallest gap between Brazil and other countries is in healthcare.
>
> *(MARKET STAKEHOLDER)*

> I think that, on a global scale, we're in the early stages when it comes to healthcare. [...] I don't think Brazil is that far behind in AI. We might still be behind in healthcare, but I believe everyone is more or less in the early stages [of applying AI in the healthcare sector].
>
> *(MARKET STAKEHOLDER)*

> I believe there are already many practical experiments happening in healthcare, in both the public and private sectors. But I don't think we have a diagnostic perspective yet regarding the inventory [of initiatives] to know where we stand in terms of maturity compared to other countries internationally. My impression is that we're not at the forefront; [...] but for a country in our situation, we're making progress in a more general context.
>
> *(PUBLIC SECTOR STAKEHOLDER)*

Regarding the Brazilian case, the general perception is that the development of AI applications has taken place in a fragmented way, i.e., by way of specific actions developed by different players, without any articulation. This fragmentation can hinder the benefits of AI and limit its positive impact. This may be related to the lack of a national strategy, the diversity of the actors involved (who focus on individual experiences), and the absence of governance of large healthcare databases. Efficient data integration and the coordination of initiatives are essential for maximizing the benefits of AI in healthcare. The interviewees understand that many projects are currently being developed in Brazil in universities, private companies, and even the public administration sector, but they face difficulties in terms of practical implementation. When this happens, the application is limited to specific contexts and lacks scale and reach; they are generally localized and isolated initiatives. The transition from the design and prototype phase to the practical application phase is even more nascent with AI tools that support clinical practice; in this sense, those

solutions that support management are more easily disseminated. The interviewees believe that the development of this agenda in Brazil takes place via a bottom-up dynamic based on localized practices in some institutions and without any general guidelines. Even in organizational contexts (such as corporations and public and private healthcare facilities), these movements are led by one or just a few groups of professionals rather than by all-embracing institutional policies.

I'd say that there's a very high expectation from the market that AI can be a "silver bullet" for delivering healthcare, whether that's primary care, specialized care, or in the pharmaceutical sector. People have a lot of expectations, but in practice, I see very little action. I see that actually delivering AI in practice is difficult.

(MARKET STAKEHOLDER)

At this very early stage, we're still in the phase of organizing and gathering data and thinking about how to interoperate. If we're going to talk about large-scale, I believe we have some very specific cases.

(HEALTHCARE FACILITY STAKEHOLDER)

Often, it's the startups and the doctors who are activating this; it's still very bottom-up, isn't it? Few companies... few health companies have said: "Look, from now on, the top-down approach is going to be like this."

(MARKET STAKEHOLDER)

I think that some things are already very well-structured and functional, but there aren't a lot of initiatives. [...] AI is being used a lot in imaging, and there's a lot of AI ready for use in management, but not many people use it. I think that now that generative AI is coming into play, people want to use it in the customer service area.

(ACADEMIA STAKEHOLDER)

I think there are islands of excellence, and some quite robust initiatives are currently being used in the private sector. [...] In short, there are several one-off initiatives that are not very similar but that have a certain synergy on the subject.

(MARKET STAKEHOLDER)

Among those interviewed, there is a common understanding that the market — corporations, startup*s*, and health insurance companies — often working in partnership with academia and research centers, has been leading the recent movement to develop and disseminate AI solutions in healthcare in Brazil. Stakeholders from different segments also indicate that the public sector is lagging behind in this race: The topic is just beginning to enter the public agenda and has its great potential recognized. According to a significant proportion of those interviewed, however, the development of this agenda in

the public sector is likely to take place at a slower pace due to bureaucratic and regulatory issues and a lack of resources. The interviews reveal that initiatives in the public sector to develop and adopt AI in healthcare also take place in a decentralized and territorialized way, without scale or specific guidelines on how to use it: There is no organized and centralized national strategy to encourage the use of AI in healthcare.

Despite this gap, which was stressed in the interviews, there is a strong perception that there is great potential for using AI in the Brazilian public health system, thanks to the Unified Health System (SUS) and its (nationwide) databases. Considering the reach and scale of this system and the diversity of the Brazilian population, the data collected by the SUS are of enormous value, and, if they are processed and prepared appropriately, this will represent an essential step towards achieving data integration.

Data integration was mentioned in most of the interviews as a crucial point for developing AI tools. The subject was presented as a key element in both the opportunities and challenges for developing AI in healthcare in the country, which will be addressed in the following topics. For now, it is essential to reiterate the perception that the Brazilian public health system and its information bases are strategic for data integration at the national level.

The following quotes illustrate the current picture of AI as applied to healthcare in the Brazilian public sector. The words of the stakeholder*s* highlight the government's challenges when it comes to absorbing technological innovations, according to the perceptions of people from different segments. Various reasons are cited, such as a lack of resources and bureaucratic issues. The quotes also reveal no coordinated incentives or clear guidelines for applying AI tools in government healthcare policies. In short, the interviews show that, despite the SUS's great potential for linking initiatives and integrating data, the inclusion of AI in public health today faces significant fragmentation and a lack of articulation.

One of the main elements is the huge inequality — even iniquity — between the public and private sectors. The private sector is already mature in terms of algorithms for analyzing the population's health and identifying the risks and vulnerabilities associated with working with the population. They are being used, for example, by some health insurance companies and large corporations and are having an incredible impact; using these large databases and electronic medical records, for example, has a huge impact. But in Brazil, we have the SUS's database, which is one of the largest collections of medical records in the world, if not the largest. But there's a lack of solutions, and the regulatory system itself is lacking... We face great challenges like [developing] a national policy for specialized care.

(PUBLIC SECTOR STAKEHOLDER)

The public sector ends up taking a little longer to absorb these changes and these technologies, and the problem is not a lack of knowledge. It's more a question of bureaucracy, of what's allowed and what's not. When you're in the private sector, you have a bit more freedom to explore what's new. But I see that although the public sector may be a little more cautious, it's not letting it go; it's developing projects, and it's looking for information

(PUBLIC SECTOR STAKEHOLDER)

Everything you have today in relation to AI in the public sector is much more momentary, a political issue connected to the mayor or the state or municipal health secretaries rather than being part of a strategy or a government plan. [...] Currently, there's a disconnect between the Ministry of Health (MS) and the states and municipalities in relation to the use of AI. There's no clear policy today. We don't have a clear idea as to why a mayor should incorporate AI into the Family Health Program to help the community health agent because there's a lack of digital literacy.

(MARKET STAKEHOLDER)

We'd need better investment incentives from the public sector. I'm talking specifically about the SUS. For us to develop solutions in this field that have a more global impact, we must work more on network cooperation between municipalities, states, and at the national level rather than relying on localized experiments that we're going to scale up later. After all, the SUS is about cooperation, isn't it?

(PUBLIC SECTOR STAKEHOLDER)

Brazil doesn't have a national AI project, unlike the vast majority of other countries of the same size that have similar economies. [...] We still don't have projects to encourage it. There's no national strategy for AI.

(ACADEMIA STAKEHOLDER)

In [public] management, we're still getting into the analytics layer. We're still setting up what would have been the old circulation or command rooms, and moving towards BI technologies, Big Data. It's still very little.[...] And on the clinical side, we're starting to see something that has a decision support system, but it's still in its early stages.

(PUBLIC SECTOR STAKEHOLDER)

I believe that if we look at the private healthcare environment today, we're doing and using things that are comparable to what's being done and used in countries like the United States and in Europe. We have niches of excellence within the SUS, but if we look at the overall volume [of things being done], there's still a lot missing.

(ACADEMIA STAKEHOLDER)

The interviewees also realize that AI is not evolving evenly across the different sectors and areas of healthcare. As mentioned, using AI to manage and administrate healthcare facilities is more advanced than in clinical practice. In those medical areas with the most significant progress in the development and application of AI in healthcare nationally, radiology stands out. The use of AI in imaging and diagnostics is better consolidated and incorporated into everyday practices, not least because, in the view of those we interviewed, healthcare devices are starting to include this type of technology, which is why university and specialization courses are preparing professionals to work with these tools. Below are some statements that highlight the progress made in applying AI in the field of radiology in Brazil:

It's more functional in imaging. It's already possible to use it in this area.

(ACADEMIA STAKEHOLDER)

Of course, some areas are more advanced than others, such as radiology. I see a lot of things that already work, but they're also embedded in the software and devices, aren't they?

(HEALTHCARE FACILITY STAKEHOLDER)

It's important to point out that many projects obviously revolve around radiology because even in the US, 75% of the algorithms that have been approved by the Food and Drug Administration (FDA) are in some way related to radiology. So, it's radiology. Because diagnostics is a technological area, which was where AI was more widely used from the outset.

(HEALTHCARE FACILITY STAKEHOLDER)

Some areas are more developed than others, so in healthcare, I believe that radiology is an area that's going to benefit from the contribution of AI. So, I understand that radiology is the main area that's benefiting. Then you also have those areas or sectors that are interested in electronic medical records, which can also benefit.

(HEALTHCARE FACILITY STAKEHOLDER)

When discussing the progress in AI as applied to healthcare in Brazil, some interviewees identified the growing demand from healthcare professionals for knowledge and qualifications related to the topic. Some cited the substantial increase in interest from different healthcare professionals, including students and newly qualified doctors. The perception that there is more dialogue between information technology (IT) and medical professionals in building, improving, and maintaining AI tools applied to the healthcare context was also mentioned. This is

considered positive and means, in the view of the interviewees, that progress is being made in putting together multidisciplinary, qualified teams to build and operate these tools.

The perception I have — not least because I'm in medical school — is that medical professionals are very interested in it. So, today, we already have radiologists who are learning how to work with these algorithms and even developing them. Nowadays, a good part of the implementation has already been done, so we end up doing a lot more customization, like adjusting parameters. Even so, you must have a certain programming command: You need to know how to work in that environment. We have young doctors who are graduating and already have an interest and familiarity with computer issues and who are starting to get to grips with it and moving forward.

(ACADEMIA STAKEHOLDER)

On the one hand, there has been growing interest in recent years from the exact sciences: From the computer engineering and physics communities. On the other hand, health professionals themselves — doctors and other professionals, such as nurses — tend to be very active in the area of digital health. We see a rapprochement between the two sides. I see professionals in computer science and engineering increasingly collaborating with the healthcare sector because of its significant impact. On the other hand, healthcare personnel are also beginning to work with these models and understand them.

(ACADEMIA STAKEHOLDER)

Still considering this overview of the current state of AI in healthcare in the country, the interviews revealed gaps and weaknesses. The aspects identified by the interviewees as the central gap to overcome in this early stage of AI technologies in healthcare in Brazil relate mainly to the availability, governance, and regulation of the data.

Regarding the topic of data, which is explored in depth in other sections of this chapter, the study's interviewees believe that the lack of quality data and/or the difficulty of integrating large databases are hindering the progress of AI in healthcare in Brazil. The lack of an adequate architecture for making data available and using it is considered a significant national weakness.

So, we need to be a little more careful, keep up with the changes, and prepare for when Brazil is really ready to include algorithms in clinical practice, which will depend on the quality of the data. It's going to depend on overcoming some of the major challenges of including algorithms in clinical practice, which is, first and foremost, the question of the quality of these algorithms and their performance in each specific location. We'll see if they're working in the different Brazilian realities.

(ACADEMIA STAKEHOLDER)

The biggest issue in hospitals today is the quality of the data. For me to use any AI product, I need to have a top-quality data layer because otherwise, my product won't be worth anything; it won't reach its full potential.

(MARKET STAKEHOLDER)

Regarding regulation, there is a widespread perception of a significant gap in the current Brazilian context. Interviewees from different sectors believe there is a lack of regulations and guidelines for developing and applying AI tools in healthcare. They also identify a lack of governance structures for monitoring, validating, and supervising these applications in practice.

> We see solutions being implemented more quickly [in Brazil]. The environment is very fertile, and a lot of solutions are being implemented and developed. But, on the other hand, my concern is that this opens up another angle that I think is complicated and delicate. Because you can prove the concept faster, you can put it into practice faster as an effective clinical tool. I think this is a point that we still have a certain difficulty with because there's a lack of organization and perhaps governance, as well as maturity on the part of those who do it. It's one thing being in a lab or even in a clinical environment, working with research strategies. But it's something else to go into a clinical environment with a tool that can be used and produce [results?], but that requires maintenance and monitoring. I believe this is still not very well understood.
>
> *(ACADEMIA STAKEHOLDER)*

> There's a lack of regulation at the national level to encourage this because some of these technologies and some of these AI tools or platforms need to be approved by the Brazilian Health Regulatory Agency (Anvisa), and we have an agency that doesn't have the slightest capacity — and that's the truth — to evaluate or come up with a policy that has clear criteria for approving AI tools. So, the regulatory framework is still not corresponding to this.
>
> *(MARKET STAKEHOLDER)*

The interviewees generally believe Brazil needs to address these two major challenges, specifically: the architecture for data availability, and governance and regulation. This needs to be done to overcome the current stage of development and prepare for the potentialities and risks that AI brings to healthcare. Infrastructure and cultural problems, such as resistance to innovation, were also mentioned but less frequently and were less relevant to the interviewees' statements.

BOX HIGHLIGHTS - CURRENT STATE OF AI IN HEALTHCARE

- There is a prevalence of very positive expectations about the benefits of AI to Brazil's public and private health systems. Prospects for reducing costs and increasing the efficiency and scalability of health services.
- A general perception that there is vast potential for using AI in healthcare in Brazil, both in management and medical care. There are specific areas at a more advanced stage of development, such as radiology.
- There is an increased interest and enthusiasm on the part of healthcare professionals and students due to the optimistic scenario and growing interaction between IT and medical professionals in the construction and operation of AI tools in healthcare.
- A consensus is that the country is at an early and experimental stage in using AI in healthcare. Solutions are developed in a fragmented way, with little articulation.
- A perception that initiatives generally originate in the market due to different constraints specific to public organizations, such as greater bureaucracy, a lack of resources, and the need for greater zeal in public practice.
- It is challenging to move from the prototype phase to practical implementation in relation to operationalization, especially in the case of tools to support clinical practice, which depends on the training of professionals and the dynamics of the hospital routine.
- A lack of regulatory standards and guidelines for developing and applying AI tools in healthcare is pointed out as a significant gap and a lack of governance structures to validate and oversee the use of AI tools in the sector.
- A lack of architectures for providing and integrating quality data.

OPPORTUNITIES FOR AI IN HEALTHCARE

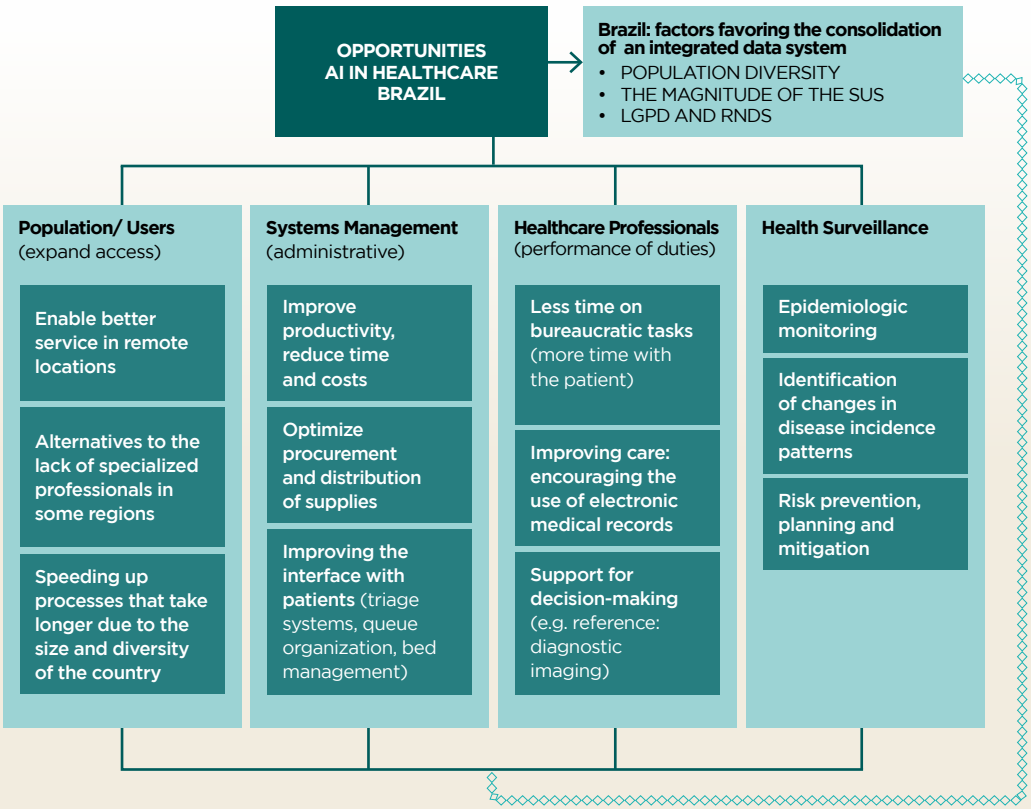
This subsection is dedicated to presenting the stakeholders' perceptions of the opportunities for AI applied to healthcare, considering the specificities of the Brazilian context. The interviews sought to capture views on AI's potentialities: Which sectors, areas of healthcare, types of tools, and applications have the most significant potential and represent the best opportunities for the Brazilian reality. The study also aimed to understand the bottlenecks in the country's healthcare system that this technology could help address.

As mentioned, there is a lot of optimism among the study's interviewees, who see many opportunities for using AI in healthcare in Brazil. The potential in various sectors and processes of healthcare management and clinical care was

pointed out, with improvements envisioned in different institutions and for the stakeholders involved in this ecosystem, and benefits for the community as a whole. Figure 2 provides a summarized overview of the opportunities we identified in the interviews.

> I believe that [with AI], it's possible to expand the range of services offered, scale up operations, improve access, and enhance diagnostic accuracy. AI provides valuable information for decision-making, both at the level of clinical care — where AI has proved to be superior to human sight in imaging exams — and in predicting future public health emergencies, for example. In the areas of surveillance and prediction, we also have significant potential, and Brazil is already starting to benefit from this impact.
>
> *(PUBLIC SECTOR STAKEHOLDER)*

FIGURE 2 - OPPORTUNITIES FOR AI IN HEALTHCARE IN BRAZIL



SOURCE: PREPARED BY THE AUTHORS.

Population/users

Considering Brazil's particularities and recognizing widespread inequality in access to quality healthcare, many interviewees identified AI as a technology capable of mitigating the effects of this particular social problem. Given the significant territorial disparity in the healthcare system, which results in a shortage of services, facilities, and professionals in many regions, AI tools could connect the patients and professionals in these areas to specialized knowledge for supporting operational and clinical decision-making.

For example, AI solutions can help available professionals predict clinical outcomes and speed up referrals in scenarios with a shortage of specialist doctors. The interviewees consider medical decision support systems to be good opportunities for reducing inequality in access to qualified care and streamlining flows and processes. This progress is particularly relevant in Brazil, where the inadequate distribution of resources and professionals results in a lack of services and slowness in health care, diagnosis, and treatment.

One thing is certain: In the same way that algorithms have transformed other areas, they're also going to transform healthcare and guarantee quality care, especially in the more remote regions of Brazil where there's only one doctor, for example. There are lots of Brazilian cities where there's only one doctor; he has to make decisions in all the specialties. He doesn't have a cardiologist or a pulmonologist to refer to... he has to make all these decisions. With the arrival of these algorithms, he'll have the help of the best doctors in the world, thus reducing the immense inequality that we have in healthcare in the different regions of Brazil. With the advancement of this AI algorithm, we're going to provide quality specialist care in the country's remote regions.

(ACADEMIA STAKEHOLDER)

Another opportunity is that with AI, you can offer services to the population that are currently very expensive. For example, let's consider a patient with a heart condition or an elderly person with diabetes living alone. [...] With AI and cloud technology, you can monitor this patient who, today, is at home and sometimes may not have the money for transportation or may live far away. So, you could provide the same care from a doctor who's here [in Hospital Center A][6], who could, for instance, automate certain clinical protocols. José lives in the Amazon and has heart disease or diabetes, and the nearest health center is 100 km from his home. He could answer the same questionnaire and fill the same forms, and the same bot that serves São Paulo would serve the guy who lives there [in the Amazon]. If AI detects that João is in the middle of the Amazon and his nearest health center is 100 km away, the doctor will be able to know that this guy needs to be seen by a health professional.

(MARKET STAKEHOLDER)

6 Hospital center with headquarters in São Paulo and a branch in Brasília. Considered a benchmark in healthcare in Latin America, it is part of a network of healthcare specialists and training programs.

We know that there are several "Brazils" in Brazil. So, if you think that in the North, in the Northeast, or even here in the state of São Paulo, you need a qualified doctor, a decision support system, and triage... There are opportunities in these areas for supporting places where there's a shortage of doctors; a radiologist, for example, or even a pathologist. [...] A decision support system would, then, certainly provide more accurate support. I can envision opportunities both in screening and in decision support for places that need greater expertise. So, imagine a trained system with millions of images and a resident physician, a doctor, who's not a specialist in the area, [but is living] in a place like this... I believe that would make a very valuable contribution.

(HEALTHCARE FACILITY STAKEHOLDER)

Systems management

Regarding administrative and operational issues, interviewees widely highlighted AI's potential to increase productivity and efficiency in the management, operation, and logistics of healthcare systems, services, and facilities. They particularly noted improvements in workflows and processes, reductions in time, and optimizing financial and human resources, with potential benefits for public and private healthcare systems.

There's a component for improving financial workflows, agreements, purchase requests... So, I think all these administrative tasks tend to evolve [with AI].

(HEALTHCARE FACILITY STAKEHOLDER)

Reducing costs. The moment I make a prediction, a forecast, that a citizen may need high-cost health care and address this in primary care, these are resources [saved], which benefits citizens.

(PUBLIC SECTOR STAKEHOLDER)

The interviewees were optimistic about using AI tools to organize the logistics and internal work processes, such as purchasing and distributing supplies, equipment, and medicines and allocating human resources. Expectations are also high regarding improvements in workflows and operations that interface with patients, such as triage, queue organization, and the prioritization of care and in-bed management and scheduling examinations.

Many of the current apps focus on management, and they're also very important: Optimizing the flow of care, equipment use, and bed occupancy. I'm optimistic about the potential of these tools.

(ACADEMIA STAKEHOLDER)

AI's impact on operational efficiency. Speeding up the use of operating rooms and algorithms for accelerating MRI scans end up reducing the exam queue and improving room and bed turnover and patient safety within the hospital.

(HEALTHCARE FACILITY STAKEHOLDER)

> Another important area we have, which is found in all hospitals, is queue management. Managing surgical or procedure queues, for example: determining the criteria for prioritizing these queues. We've been using automated methods based on patients' care histories to suggest reordering queues when the criteria are already well established. We're also employing machine-learning methods, specifically AI-driven approaches, for this purpose.
>
> *(ACADEMIA STAKEHOLDER)*

> So, workflows, identifying priorities, and managing queues for critical cases. I believe these are areas with significant room for improvement [in Brazil]. For instance, the AI tool I mentioned: If I have an emergency room that performs a lot of tests, if I have AI that issues an alert to the team every time a very serious case appears... You're not going to be relying on a process whereby a human takes a look and identifies that it's serious before being able to act. This would save a huge amount of time. I believe that predictive models, in general, have enormous potential for providing benefits.
>
> *(HEALTHCARE FACILITY STAKEHOLDER).*

Improving these management aspects can optimize healthcare systems as a whole and increase their capacity to absorb demand and provide care to expand the population's access to healthcare. In this sense, the interviewees believe that improving the operational flows of healthcare systems and facilities is one of the main potentialities of AI in Brazil.

The interviews indicated that these AI applications in management and operations are already a reality in some contexts, especially in the private healthcare system, where their positive effects are already seen in practice.

Health professionals

As mentioned, the interviews also highlight significant potential for AI in clinical and care settings. A substantial portion of the stakeholders noted the potential of AI tools for supporting the daily work of frontline healthcare professionals, such as doctors, nurses, and physiotherapists. The interviewees pointed out that these tools could reduce the time spent by professionals on bureaucratic and administrative tasks, such as filling out forms, documents, and medical records. With these resources in place, professionals can spend more time listening, paying attention, and offering human-centered care. Optimizing the working time of healthcare professionals will also boost productivity in patient care, improve their daily work environment, and make clinical practice more effective and focused, ultimately benefiting patient care.

This greatly increases the productivity of doctors and reduces the time they waste dealing with bureaucracy. Much of a doctor's time today is spent filling out paperwork. Algorithms can do this automatically via the doctor's conversation and by identifying the procedures that have been carried out. This will increase the potential for doctors to use their time in the most productive way and in the way they would like to use it, which is seeing patients and not handling internal bureaucracy.

(ACADEMIA STAKEHOLDER)

Let's look at the [healthcare] professional. [...] When we look at the work of healthcare professionals today, and there are several articles about this, it's impressive that more than 30% of their time is spent behind a computer filling out paperwork. [...] So, I think the first point is that AI helps a lot in this process of filling out paperwork.

(HEALTHCARE FACILITY STAKEHOLDER)

The interviewees widely mentioned the potential of tools related to medical records because of their potential to improve efficiency when accessing and managing patient information. Also mentioned were apps that automatically record data in the electronic medical record system by transcribing spoken information from consultations and other situations, enhancing and synthesizing critical information from the patient's history. The interviews also highlight the advantages of electronic medical record systems that integrate and organize valuable patient life history information. The development of these applications also has the potential for producing more in-depth analyses of the current clinical state and providing visualizations of patients' future conditions through early prognosis and diagnosis. According to the interviews, AI tools for recording, optimizing, and visualizing medical records would also result in more productive and efficient care workflows, which would benefit healthcare professionals and provide more effective patient care.

Nowadays, the healthcare team doesn't have time to look after a patient; there's just no time. You see an electronic record that's full of laborious information, with 50 previous examinations. The doctor doesn't have half an hour or an hour to spend looking at all that. He has just five minutes to read the history and ten [minutes] to deal with the patient. So, how do I make sure that in those five minutes, he can digest all the data that already exists? The system can digest all this mass of data, refine it more than mine it, and generate an ultra-qualified lead of the people who need to be seen right away.

(MARKET STAKEHOLDER)

I see [an opportunity] in medical records. [...] Burnout [among professionals] associated with the use of medical records is one of the main problems in healthcare. [With AI] you can focus on delivering care to the person rather than completing medical records. The second thing is how we analyze these records. [...] So, we can simultaneously transform the records and analyze them better, and I can generate different visualizations for each user based on their profile. I can structure totally unstructured data, but I can also give it back to the person; I can give it back to that person so we can take better care of them.

(MARKET STAKEHOLDER)

Another thing I see [as potential] is that you can provide healthcare professionals with advice, which is another big field. [...] Summarizing the person's use history and their history of illnesses, which in medicine we call previous pathological history, speeds up the doctor's decision-making process. An example: An elderly patient who had a stroke four years ago was discharged and treated at a rehabilitation clinic and was classified as fit after rehabilitation. So, if I give the professional a summary like this, I help them decide quickly about the problems.

(MARKET STAKEHOLDER)

The interviewees also mentioned many potentialities of applying AI to support decision-making in clinical practice, such as tools that can help formulate diagnoses and systems based on predictive models for visualizing future clinical outcomes and prevention strategies.

The main opportunity in the healthcare context is supporting medical decisions.

(HEALTHCARE FACILITY STAKEHOLDER)

The big expectation is that we'll start having personal assistants to help the doctor. So, helping the doctor to end a consultation, transcribing things that the doctor says during the process into the electronic systems, comparing any patterns, saying that it's not right, and suggesting therapies; that kind of thing. There's a big field [of opportunities] that we don't know everything about yet. [...] You'd see more patients, and therefore have more access. By improving quality you'd also be able to go into areas that don't yet have it, so you'd be improving the quality of care provided.

(ACADEMIA STAKEHOLDER)

There's an application [of AI] that's very direct with the professional. You have an algorithm or a model that helps diagnose and predict, for example, the evolution. It helps predict clinical outcomes. So, based on what we have in the medical records and a model that's been previously trained with a large set of data, we can have a reference for the clinical outcome over the next six months, a year, or two years. This is phenomenal. [...] Thinking about the public health of the community, you can predict the evolution of populations with more likely outcomes. You can predict trends in a population that's aging... I've seen work with the elderly to predict the tendency of elderly people to fall. These direct applications for healthcare professionals can improve care.

(ACADEMIA STAKEHOLDER)

On the clinical side, and thinking about early diagnosis and prevention, I think we're going to work a lot with AI in terms of preventing the patient, the individual... even before they're a patient, having their information collected somewhere and obtaining insights from them, so they can improve their health care and don't need to be hospitalized.

(HEALTHCARE FACILITY STAKEHOLDER)

AI tools applied to imaging exams are indicated as good examples of solutions that support decision-making and reduce the time involved in reports and diagnoses, as expressed in the following quotes:

Another example is diagnostic imaging. Today, we're already seeing various imaging diagnoses that use AI to speed things up. You do a CT or MRI scan, and it sometimes takes ten or fifteen days to get the report back. With AI, the report is immediate, but obviously, a human — a doctor — needs to check it, and if there's any difference between what the AI has pointed out and what the doctor thinks, another opinion must be sought, but this also speeds things up. Instead of having [to wait for] a report that's going to take a week or more, this report is immediate.

(PUBLIC SECTOR STAKEHOLDER)

If I were to choose ground that's fertile in terms of opportunities, where we can achieve quick results that have a wide impact, it would be imaging. It's particularly good that we have an algorithm for identifying X-ray images so we can issue a report more quickly. This is very important, and a lot of progress is being made in this regard.

(PUBLIC SECTOR STAKEHOLDER)

Health surveillance

Concerning public health and health surveillance, AI's potential to assist in epidemiological monitoring and identifying changes in disease incidence patterns in the population was mentioned by interviewees. AI tools for such purposes can improve the ability to prevent, plan for, and contain health risks and emergencies.

So, thinking at the SUS level... You have a tool that's not very sophisticated, but that helps you organize patient flows and identify certain patterns in the population that will help you screen and stratify certain types of cancer, for example, or chronic diseases. [...] So, the possibility of having tools that will help in this process of identifying the most critical things within a very large context... these tools are especially useful. [...] If you're in management and have to make decisions... there's always a shortage of money, and the population is very big. So, where do we go from here? [...] The possibilities [opportunities] are enormous.

(ACADEMIA STAKEHOLDER)

It's [the potential of] the population vision, a vision that we can reach more people in a shorter time.

(PUBLIC SECTOR STAKEHOLDER)

[...] It's the opportunity to detect outbreaks and epidemics of pathogens with epidemic or pandemic potential. I don't think it's going to be the first time you hear this today, but Brazil produces data from everywhere. The Ministry of Health has more than eight hundred different databases on aspects ranging from clinics and laboratories to logistics and supplies, the vast majority of which don't communicate with each other. We have a wealth of information in this data that allows us to detect, for example, the start of an epidemic or a pathogen, even before the official surveillance systems can do so [...].

(ACADEMIA STAKEHOLDER)

Favorable factors for consolidating an integrated data system in Brazil

Various interviewees pointed out that Brazil presents a particularly favorable scenario for the use of AI in healthcare due to the large volume of data, which would enable the construction of tools based on a robust and diverse mass of information. This perception is linked to the country's large population and extensive territory, our characteristic genomic diversity, and a single public health system that serves many users and systematically records their information. These characteristics of Brazil are favorable to the development of consistent machine-learning algorithms. Furthermore, as mentioned in the previous topic, the health data systems and repositories of the SUS are assessed as having enormous potential for integration with a view to interoperability.

Brazil is a country with 210 million inhabitants and is the only one with more than one hundred million people in an underfunded but structured single health system. No one else has that. We're a laboratory for the world. It's not just because of our genetics; it's because we have this system of continental dimensions. [...] These tools have to be integrated, so integration is an opportunity for us.

(HEALTHCARE FACILITY STAKEHOLDER)

Healthcare [in Brazil] is extraordinarily rich in data: data on citizens' lives both in the SUS and in complementary healthcare, which is private. With the right research and the right perspective, which is not just that of IT but of the healthcare professional, this data can do a lot for citizens. They can anticipate health situations, avoid major problems, and reduce costs. [...] So, there are many cases we can use in healthcare, very much in the sense of predicting what a citizen might have so they can avoid it.

(PUBLIC SECTOR STAKEHOLDER)

> One advantage we have is the SUS, our unified [healthcare] system, which is very large and has a lot of data. So, I think that compared to other countries, we have the potential to have a health database in the SUS that is potentially enormous: One of the largest in the world in terms of data quality and quantity. I know it's making a lot of progress, but I think there are some strategic initiatives to organize the SUS and DATASUS data that are making a lot of progress. So, this is an opportunity. [...] The SUS has the potential to see everything, from care data to administrative data, but we still have a lot of ground to cover to make this data actionable and usable.
>
> *(HEALTHCARE FACILITY STAKEHOLDER)*

> [Our] health system is the biggest generator of data in the world, but we don't use it. So, there's no need to generate new data. Just take the data that's s already there and transform it.
>
> *(MARKET STAKEHOLDER)*

Still, about the data, the interviewees pointed out two other aspects as being particular strengths of the reality in Brazil, albeit they did so less frequently and emphatically. These are the National Health Data Network (RNDS) (MS, n.d.) and the General Data Protection Law (LGPD) (Law No. 13.709/2018).

BOX 1 - NATIONAL HEALTH DATA NETWORK

The National Health Data Network (Rede Nacional de Dados em Saúde [RNDS]) is the Brazilian healthcare interoperability platform that was set up in 2020 to promote the exchange of information between points in the healthcare network, thus enabling the transition of care and its continuity in the public and private sectors. Its constitution is a structuring part of Conecte SUS, a program aimed at the digital transformation of healthcare in the country; it is connected to the DHS. More details on the network's functionalities and potential benefits can be found on the MS's RNDS website.

RNDS

BOX 2 - GENERAL DATA PROTECTION LAW

Established by Law No. 13,709 of August 14, 2018, the General Data Protection Law (LGPD) sets out the guidelines for processing personal data in Brazil, including digital media. It aims to protect the freedom and confidentiality rights of citizens and institutions.

LGPD

The interviewees pointed out that the LGPD was a regulatory advance that would enable the proper development and application of AI tools in Brazil. With this initial regulatory framework in place, it is believed that applications can begin to be developed, evaluated, and applied. In national contexts where this type of regulation on the use of data has not yet been put in place, the evolution of AI would be hampered.

The RNDS, on the other hand, was seen as a positive strategy with the potential to promote data interoperability in the future. The prospect of data and systems integration on the horizon, as put forward by the network, can be seen as the country's potential for the evolution of AI applied to healthcare.

Looking ahead, we're in a very promising scenario [...] because we're finally bringing in the area of knowledge of healthcare informatics. [...] Today, we have the prospect of the RNDS and interoperability between the public and private sectors and between all spheres of management.

(PUBLIC SECTOR STAKEHOLDER)

In addition to these potentialities specific to the Brazilian context, the interviews gathered various perceptions about AI's potential as applied to healthcare in general. Many comments emphasized how AI as a technology can help solve problems and improve multiple aspects of healthcare in the country that are not limited to the reality in Brazil. In general, these statements point to the significant potential of AI in administration and operational aspects, on the one hand, and in clinical practice and patient care, on the other.

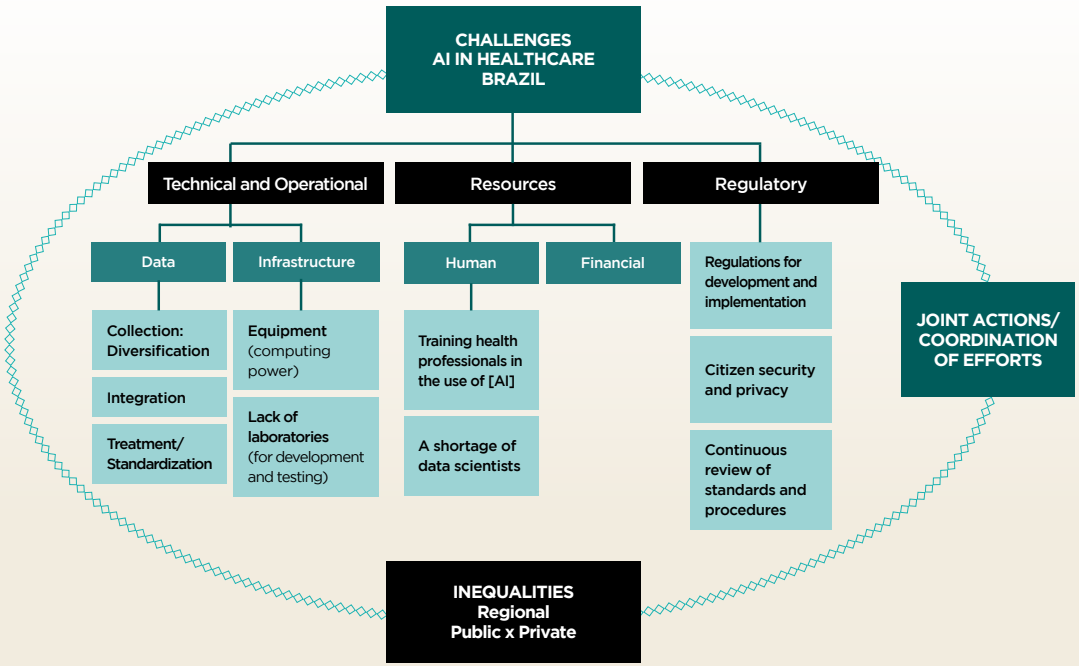
BOX HIGHLIGHTS - OPPORTUNITIES FOR AI IN HEALTHCARE

- Optimism about the use of AI in Brazilian healthcare: Identifying opportunities in various sectors and in management and clinical care processes.
- The expectation is that AI can minimize the territorial inequalities in access to quality healthcare that are typical of the context in Brazil.
- Opportunities for the patient: Expanding the supply of health services, improving access and diagnostic accuracy, and connecting patients to specialized knowledge, especially in remote regions.
- Opportunities for healthcare providers: Optimizing internal workflows; improving bed management, waiting lists, and hospital logistics; and speeding up imaging test reports.
- Opportunities for healthcare professionals: Reduced time spent on bureaucratic tasks, AI tools for analyzing and interpreting medical records, and support for medical decision-making.
- Opportunities for public health and health surveillance: Epidemiological monitoring; identification of changes in disease incidence patterns; and prevention, planning, and risk containment.
- Unique characteristics of the Brazilian context that make it rich in opportunities for AI in healthcare: A large volume of available data; the existence of integrated public health systems; existing data protection regulations (LGPD); and potential data integration via the RNDS.

CHALLENGES FOR THE DEVELOPMENT OF AI IN HEALTHCARE

This section will present the stakeholders' perceptions of the bottlenecks, challenges, and barriers to the development of AI in the healthcare sector in Brazil. We also seek to explore Brazil's specificities around this issue and compare them with those of other countries and contexts. Also discussed will be the challenges relating to technical and operational aspects (data and infrastructure), human and financial resources, regulatory issues, and coordination actions, which were the most mentioned in the interviews as being the main challenges for the country. By analyzing the challenges, we can identify priority areas for intervention and the development of strategies that can accelerate the adoption of AI technologies that are adapted to the particularities and needs of the Brazilian healthcare system. Figure 3 gives an overview of the challenges we identified.

FIGURE 3 - CHALLENGES FOR THE DEVELOPMENT OF AI IN HEALTHCARE IN BRAZIL



SOURCE: PREPARED BY THE AUTHORS.

Technical and operational

According to interviewees with different profiles, and as mentioned in the previous topics, one of the main barriers to the advancement of AI in healthcare is the lack of quality data and/or the difficulty of integrating large databases: Data is essential for the practical training of algorithms that can provide AI tools that perform well. Over and above volume, data quality is imperative in this case.

The interviewees said Brazil lacks a systematized collection procedure that provides structured and integrated data. For this to happen, well-defined collection and storage protocols must be established considering Brazil's territorial dimensions.

Our biggest bottleneck is having the data to train these algorithms.

(ACADEMIA STAKEHOLDER)

With open data, I know that the person took medicine and had a paid biochemical test, but I don't know the result of that test; it would be fantastic if I did. I don't know if the person passed away if they stopped taking the medication because they chose to, or because they died. But in reality, all of this data is interconnected. So, data integration is the first challenge.

(HEALTHCARE FACILITY STAKEHOLDER)

Another critical point about this strategic issue is the lack of uniformity and standardization in the country's health data. Stakeholders understand that the enormous diversity in information systems, including the terminologies used, makes integrating data efficiently difficult, compromising interoperability and the ability to use the data to train AI models. Despite some of the initiatives mentioned, such as the RNDS and e-SUS Basic Care (Atenção Básica), there is consensus that the effective integration of clinical data is still a significant challenge, compromising the usefulness of this data for AI applications.

As far as I'm concerned, today it's data collection and processing. That's the biggest bottleneck. It's about being able to gather a large amount of quality data because it's not enough to have the data. You have to have the data written down, let's say. It has to be very well organized in the way you want it to make predictions. Let's suppose I want to use AI to diagnose breast cancer in mammograms. It's not enough for me to go to the SUS and say: "Give me all the mammograms that are stored by the SUS." I'm just going to have a bunch of mammograms! I don't know who has cancer and who doesn't. I don't know if that image corresponds to the cancer that was type "a" in the biopsy or type "b" in the biopsy. So, in order for me to train the model, it's not just about having the images; I have to have the image and know that this is a normal image, this is an image of cancer, this was the type of cancer, this was a cancer of someone who died quickly, and this one had a very high survival rate until the prognosis. So, for me to create AI models, I need to have the data and a lot of information about the whole background of that thing, and that's not easy because, in healthcare, the data is all in silos, in little boxes. The image data is in the storage system. The patient's medical records are in another system. The medical records data is unstructured; it's written in free text. People document information in different ways. How do I transform this into a highly structured and organized table? This is by far the biggest bottleneck in healthcare when it comes to developing truly robust tools.

(HEALTHCARE FACILITY STAKEHOLDER)

The interviewees also mentioned the infrastructure issue as an obstacle to AI advancement in Brazil. Among the challenges cited was the high computing cost of the solutions, which makes them unaffordable in many cases. Looking at the country as a whole, the existing infrastructure is considered

inadequate and insufficient for supporting the new demands of these technologies.

There is also a perception of inequality here because technological resources are not evenly distributed between local contexts and the public and private sectors. A lack of specialized laboratories for developing and testing AI applications in healthcare was pointed out as a barrier, especially in Brazil's public health system. There are some centers of excellence in the private sector with laboratories that, although they stand out in terms of their infrastructure, face difficulties when implementing these technologies on a large scale.

Scarce financial, human, and technological resources are identified as a significant challenge in Brazil, even though this can be seen in other national contexts. The interviews indicate, therefore, that the successful implementation of AI in healthcare in Brazil will require coordinated efforts and investments on several fronts, including data, infrastructure, and human resources.

Resources

Another major challenge is the shortage of trained human resources with the digital skills needed to implement and use AI in healthcare. There are multiple aspects of this gap: Some interviewees mention the difficulty healthcare professionals have in familiarizing themselves with AI tools, while others highlight the complexity of attracting specialized data scientists. Some interviews addressed both the difficulties of the Brazilian reality and stressed the lack of trained professionals as a critical point in the effective adoption of AI in Brazilian healthcare.

I think we have good IT professionals, but, unfortunately, the standard has declined. The pandemic led to an inflation in IT salaries, and now you see young professionals with limited knowledge and experience earning high salaries. So, they often feel that they don't need to learn more — they believe they know enough already because they're well-paid for the time being. We're facing significant challenges here in finding professionals like data engineers and data scientists. We're not limiting our search geographically because, obviously, we can work with a lot of professionals remotely today. Even in Brazil, for example, the difficulty in finding qualified data scientists and data engineers is severe.

(PUBLIC SECTOR STAKEHOLDER)

Several interviews addressed the shortage of qualified technical labor as a significant bottleneck and emphasized that

the same professionals needed for developing AI in healthcare can also work in other sectors, such as finance, which offers better compensation. The interviewees believe, therefore, that attracting and retaining AI professionals who specialize in AI development in the healthcare area is a challenge in Brazil.

> You need computer resources and specialized people. Perhaps we have too many bottlenecks for this in the country. The faculties today – particularly math and computer science – that produce our data scientists... the number of professionals graduating, even in engineering, who have the skills to do this does not meet market demand. The financial market absorbs a good number of these people, so there's a lack of manpower able to work with these things in a critical way. Again, it's not about consuming an off-the-shelf product or technology that someone has trained abroad in; it's not simply a question of plugging it in and starting to use it in the hospital where they work. You need to validate it; you need to have people who criticize and evaluate it consistently.
>
> *(ACADEMIA STAKEHOLDER)*

It is important to note that the lack of data scientists is not exclusive to Brazil. This has been highlighted as a global problem. Demand has increased quickly, and countries have been unable to train people to meet it. However, the interviewees believe that some countries are tackling this issue strategically and systematically allocating resources to AI. Interviewees identified only isolated actions by technology research funding agencies in Brazil, such as the Financing Agency for Studies and Projects (Finep) and the Research Foundation of the State of São Paulo (Fapesp). However, these actions alone do not constitute a comprehensive national strategy.

> We don't have a national infrastructure for high-performance processing to train algorithms. The research support foundations in the states have issued calls for proposals to allow research groups to develop projects; even FINEP has released calls for proposals. But perhaps that's not enough given the demand we face. We're left with some significant bottlenecks that will certainly reduce our competitiveness, and we'll have to move quickly to catch up. I believe the lack of resources for this is a critical issue; the country doesn't have an AI strategy. Several countries, such as the Asian Tigers, have clearly defined where they're going to allocate resources for AI. They have a strategic plan, but I don't see that in our country yet, except for these calls from the research support foundation, FINEP. But these are very specific [calls], not a strategy. I think the country should have a national strategy for this, like, for example, the two major [players], the United States and China, which have very clear AI strategies.
>
> *(ACADEMIA STAKEHOLDER)*

The interviews, however, also reveal that Brazil faces particular obstacles in overcoming the problem of a shortage of skilled labor. It is worth reiterating that the challenges Brazil faces regarding data and human resources are also related to some

specific features of the reality in the country. One of them would be its territorial inequality, which would make implementing AI systems in healthcare even more complex. The interviewees perceive a substantial disparity between state and municipal contexts regarding available resources and a priority for innovation. The poorer states and municipalities face significant challenges when hiring basic healthcare professionals, such as doctors and nurses, which make it unfeasible to prioritize the costly hiring of IT professionals. The discrepancy in the salaries offered to data engineers in other economic sectors, which are more attractive than those in healthcare, illustrates this segment's challenge in attracting specialist IT professionals. This challenge is even greater in government administration and in Brazilian public health. The interviewees believe that due to the scarcity of resources, the public sector tends not to prioritize hiring specialists to implement AI-based solutions. The shortage of qualified professionals in this area, therefore, is intrinsically linked to the inequalities in the country, including salary issues and priorities in allocating public resources.

The following quote illustrates how the shortage of qualified human resources is intricately linked to inequality between municipalities. This disparity is typical in Brazil, where the unequal distribution of professionals in healthcare and other essential areas results in major differences in the quality and availability of the services offered. Smaller and more distant municipalities often face greater challenges, thus exacerbating regional disparities and limiting equal access to technological innovations and quality healthcare.

The health secretary of a municipality on the outskirts of São Paulo showed me a tool and said: "I wish I could develop things like this in my municipality, but how do I do it?" I told him: "The institute develops these for free and brings it to you." He said: "That is fantastic, but I wish I had someone who could do this locally. But when I tried to hire someone, it would cost me R$ 25,000." For R$ 25,000, I can hire five nurses. So, do I hire someone to sit in front of a computer doing these things or five nurses who will help the patients?

(ACADEMIA STAKEHOLDER)

Regulatory

In addition to the aspects presented, the issue of regulation for using AI in the healthcare sector also emerged as a recurring challenge in the interviews. The perception that there is a lack of regulations and guidelines for developing and applying

AI tools in healthcare is a concern shared by many experts and professionals in the field. The rapid evolution of AI technology and its potential impact on healthcare delivery raises ethical, legal, and safety issues that have yet to be addressed in Brazil. The absence of regulation, therefore, can result in gaps in the quality, safety, reliability, and effectiveness of AI applications in healthcare. The lack of specific guidelines can also jeopardize the consistency and reliability of these tools' results and increase the risk of algorithmic bias that can have negative social implications, a lack of transparency, and inequity in access to healthcare. These concerns reflect the ongoing need to develop and update regulatory frameworks as AI in healthcare continues to evolve, thus ensuring that its benefits are maximized ethically and responsibly.

The technology is still in its early stages, so it's going to be subject to more regulation in the future, isn't it? I think the main regulation is ensuring that data is not used against citizens. It's essential to guarantee data anonymization; these things need to be done. I believe it's important to establish rules for sending data abroad. This process needs to be better structured. You can send data, but it must be anonymized, there needs to be a contract, and you must know exactly what's being sent. There should be a minimum level of structuring.

(ACADEMIA STAKEHOLDER)

There's a movement and discussion [about AI regulation], even within government circles. So, it's not that nothing's happening — quite the opposite; there's a lot happening. But many of these discussions are occurring in isolation. So, there are groups discussing it, either within the university or other research groups: Other non-governmental institutions are engaged in these discussions. But is everyone being invited to the discussions where the decision is actually going to be made? No, they're not.

(ACADEMIA STAKEHOLDER)

I need to have a minimum level of management over what can be done with this data, especially the data generated by AI, which is new data based on the data you already have.

(MARKET STAKEHOLDER)

Still, regarding regulation, the LGPD was mentioned from different perspectives. As we noted in the previous subsection, some interviewees pointed out the law as a regulatory progress that would allow AI tools to be developed and applied in Brazil. However, a small group of interviewees, mainly from the market, consider it a concern: They believe that the LGPD is a regulatory framework that can be interpreted in various ways and limits innovation processes. The general public's

lack of knowledge of the LGPD could also lead to data-sharing resistance, which could hinder innovation in the use of AI.

> The lack of sharing health data or concerns related to it. You could do a survey, take ten people, and talk about what each of them thinks about the LGPD. Depending on the level of these people, many won't even know what it's about. Another group will know what it's about but won't care. And another group will be extremely concerned, thinking that we're in a Big Brother scenario, where some guy's going to have access to [details about] their whole life. So, you're still going to have these issues because people are not familiar with them.[...] So, personally, I'd say that the different interpretations of the LGPD are a bottleneck in innovation. They hinder the ability to accelerate and deliver more things, you know what I mean?
>
> *(MARKET STAKEHOLDER)*

In summary, according to most of the interviews, the lack of clear regulations for developing and applying AI in healthcare leads to legal and ethical uncertainty, which requires more effective governance for validating, monitoring, and inspecting these applications.

Articulation

Among the challenges mentioned by the interviewees regarding the Brazilian reality, the decentralization and the disarticulation of actions and policies on the subject stand out. The lack of comprehensive, nationwide strategies that link the actors, segments, and the public and private sectors is a significant bottleneck in Brazil. This aspect was mentioned in different moments of the interviews, which indicated that the necessary arrangements for incorporating new technologies in healthcare services pose a significant challenge in the country. Therefore, this lack of a clear strategic vision and efficient joint action can result in considerable delays in overcoming these barriers, especially in public health.

While private services can incorporate technology more agilely, public health services often face difficulties due to excessively strict procedures, a lack of resources, and complex governance and operating arrangements involving federal, state, and municipal governments. In this sense, the tripartite organization of public health implies the need for integration and articulation between different actors and levels of government for the cross-cutting adoption of innovative technologies, taking into account elements such as local priorities and political differences.

The interviewees understand that overcoming these challenges requires financial investment and the construction of a strategic approach to this issue, which must consider federal arrangements, territorial inequalities, regional differences, and the diversity of local priorities. Therefore, the effective implementation of advanced healthcare technologies in Brazil requires a broad vision that considers not only the technology itself but also the social, political, and economic context of each region of a country with continental proportions.

> How do you introduce a highly technological tool in municipalities that have critical issues, where the guy is much more concerned about the water people are drinking than he is with these new technologies? So, I see it like this: Linking the inclusion of these new technologies in this tripartite configuration [of the SUS] is a very big challenge. But that's due to our state structure, which I think is good — it's republican — it's got that independence, hasn't it? If this joint effort doesn't have a clear strategic vision for the future, it becomes a bottleneck. Issues that could be overcome quickly end up taking a lot of time, you understand? Sometimes, you see a small municipality that's a healthcare hub, where a private hospital offers the same quality of care as a private hospital in a major urban center, but the public healthcare service doesn't. The private sector can implement technology there, while the public sector struggles due to the need for a series of things, such as coordinated efforts, financial resources, and a bipartite agreement, which is an issue with the state. There are also issues involving alignment with the political parties. In short, this composition doesn't generate the best results.
>
> *(PUBLIC SECTOR STAKEHOLDER)*

This section presented the most significant interview perceptions regarding the bottlenecks, challenges, and barriers to AI development in Brazil's healthcare sector. Substantial obstacles include the lack of quality data, the shortage of specialized human resources, and the absence of clear regulations. The context in Brazil is also marked by territorial inequalities that make it challenging to apply this technology in a far-reaching and integrated way. Furthermore, the lack of a national strategic vision and joint action between the public and private sectors has slowed the incorporation of these technologies in healthcare services. Finally, the lack of regulations and guidelines for the development and application of AI in healthcare is a shared concern, reflecting the need for more robust regulatory frameworks that encourage and regulate the initiatives and applications of these tools.

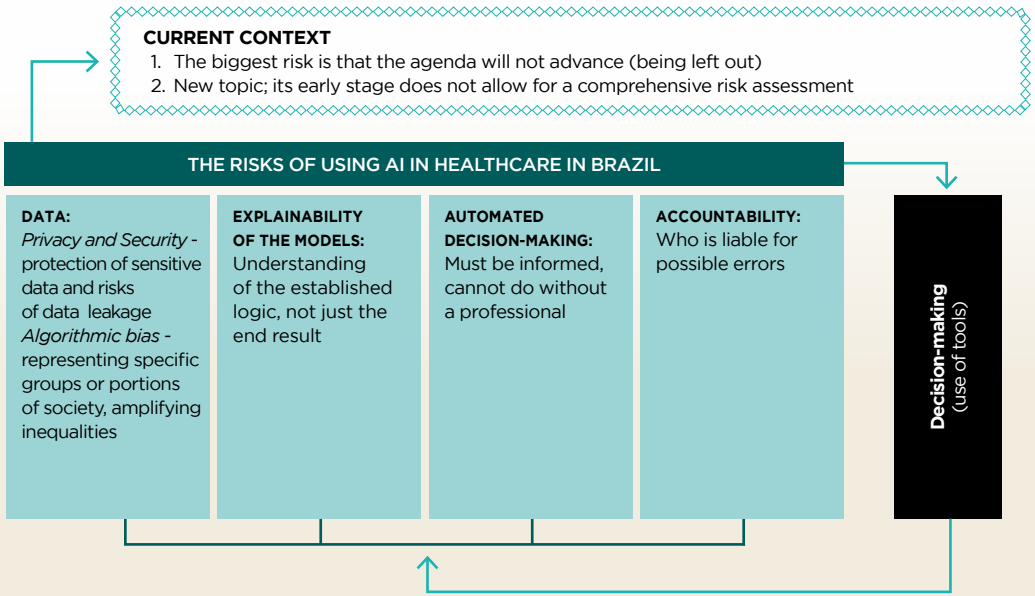
BOX HIGHLIGHTS - CHALLENGES FOR THE DEVELOPMENT OF AI IN HEALTHCARE

- Data availability and quality: This is a significant barrier to the advancement of AI in healthcare in Brazil. The lack of uniformity and standardization in healthcare data makes integration and interoperability difficult. Interviewees highlighted the need for systematized and standardized data collection, especially in more isolated regions.
- An inadequate and insufficient infrastructure. The high computing costs of AI solutions make them unaffordable. There is a lack of specialized laboratories, especially in the Brazilian public health system.
- A shortage of qualified human resources. It is challenging to familiarize healthcare professionals with AI tools. Difficulties in attracting specialist data scientists. A lack of trained professionals is critical in the effective adoption of AI in the healthcare sector.
- Territorial and resource inequalities. It is challenging to implement AI due to differences and inequalities between the state and municipal contexts. There is a disparity between private and public services when incorporating AI technologies.
- Regulation and governance. The lack of clear regulations for developing and applying AI in healthcare creates legal and ethical uncertainty and requires more active and effective governance.

RISKS OF USING AI IN HEALTHCARE

Up to this point, we have dealt with advances, gaps, opportunities, and difficulties related to the use of AI in Brazil's healthcare sector. From this section onwards, we will present the views on the possible risks associated with using AI in the healthcare sector. We sought to identify the principal risks perceived by the interviewees and the issues that, for them, should be anticipated and tackled so the country can enjoy the benefits of this technology. Figure 4 gives an overview of the elements we identified.

FIGURE 4 – THE RISKS OF USING AI IN HEALTHCARE IN BRAZIL



SOURCE: PREPARED BY THE AUTHORS.

At the outset, it is essential to note that the possible risks were not a significant theme in the interviewees' narratives. The subject only came up when explicitly stimulated but not in great depth. Among those we interviewed, there is a general understanding that the benefits of AI outweigh the possible risks associated with its use. More attention is paid, therefore, to the potential benefits than to any concerns regarding the potential risks.

This less risk-conscious view is possibly related to the early stage we are at in developing and adopting AI technologies in healthcare: This agenda is still very recent in Brazil, as some interviewees said: "We're just scratching the surface", "We're crawling." This can make it difficult to anticipate any future risks. As the following statement illustrates, the development and application of AI technologies in Brazil is taking place without an adequate mapping of the risks and, consequently, without a containment strategy for those risks.

I see that we're still just crawling when it comes to analyzing risk. It's very difficult. It's not common to find someone who makes risk-based decisions. Brazil's still crawling. There's an alignment issue, just as there is with the LGPD... From what I understand about the LGPD, there's still a need to align how much I'm going to use AI, what the patient's role is, and the impact of AI on my routine. So, I'd say that we're still crawling when it comes to risk because Brazil doesn't have a risk-based culture. [...] We don't have a scale that can guide us as to whether or not an application is a high risk for what it is intended to do. I don't have a scale. I'd say there's no path for us to follow yet. We don't have a workflow.

(MARKET STAKEHOLDER)

When asked explicitly about possible risks associated with the use of AI in healthcare, some of the interviewees again talked about barriers to the development of the technology in the country, starting from an understanding that the main risk is not moving ahead in this area, i.e., the main risk is "falling behind" in this race. They cited the risk-containment regulation as an obstacle to the development and application of AI in Brazil. To exemplify the idea that regulation can represent a risk, one of the survey respondents mentioned Bill 2338/2023, which would classify AI in healthcare as a high-risk system, regardless of how it is applied.

This definition of high-risk systems leads to a series of obligations and responsibilities for those who start using AI. In other words, it's a factor that inhibits its use because if something happens, the institution ends up being liable. So, since it's considered high-risk, this is going to lead to more time being spent developing systems and algorithms, and also to higher costs, because it's going to involve more bureaucracy, more supervision, more testing, everything... In other words, it's going to slow down the entire process of incorporating AI. And we don't really see AI as a high-risk system, do we? We see AI as a new technology that needs to be regulated, but not as a high-risk technology; on the contrary, it can bring enormous benefits to the healthcare sector.

(MARKET STAKEHOLDER)

Data: Privacy, security, and algorithmic bias.

Despite the difficulty in visualizing future risks, several points were frequently raised. One recurring concern was data security, which often arose in discussions about risks. Interviewees expressed their worries regarding data protection mechanisms and ensuring the confidentiality of sensitive information. Therefore, the risk of data breaches is highlighted as a significant issue.

The greatest risk, in fact, is the risk to privacy, the risk of individual patient data being leaked. This is particularly true when more sensitive diseases are involved, the details of which should not be made public. What if this data was leaked? But this is something that's already done today. Data is already collected, regardless of AI. These are things that have evolved independently. We've very little AI in clinical practice, but we have a huge amount of collected data from these patients. And it's this data collection that's one of the major problems we have today in healthcare, which is the risk of leaking this individualized data, which, as I mentioned, has nothing to do with AI.

(ACADEMIA STAKEHOLDER)

Some interviewees reported that although institutions generally comply with the LGPD when collecting and recording data, regulatory loopholes can weaken information security. They pointed out that AI tools that work with data collected following the LGPD can generate new data, which can be used in other contexts. There is a converging view among the interviewees that data processing is reasonably well-regulated in the country, mainly by the LGPD. Still, the availability and secondary uses of this data need clearer rules. Until this happens, there is a risk of data leakage and/or misuse.

As soon as a third-party company uses your data to generate new data, whose data is it? I know that patient data is well regulated, but who does this new data, which has been generated from the patient [using the AI tool], belong to whom? So, we've not yet seen any structure to guide us on this.

(MARKET STAKEHOLDER)

It's better to have a restrictive rule that's clear than no rule at all. [...] With a [AI] solution, I can generate data and offer value, and that data can be useful for other solutions. But who guarantees that I'm going to use it in the best possible way? So, should this process be carried out by the hospital or by my client? I don't think it should, because every client will have a different process, and a different way of managing this data. So, can we at least have a regular basis that my own clients can base themselves on?

(MARKET STAKEHOLDER)

The use of databases, this frantic search for databases, can often involve very sensitive ethical issues. For example, how long will this data be stored and reused? How will it be reused? Why is it going to be reused? You can change the image, you can anonymize it, and then you cross-reference it so much that you end up "de-anonymizing" it, and someone can be identified. And then you can have access to sensitive information, which may be private. After their death, this information can change a person's image, let's say. How do we deal with this ethically?

(PUBLIC SECTOR STAKEHOLDER)

According to the interviews, developing algorithms also involves risks that must be mitigated. Although the development of algorithms needs to be robust, and based on huge amounts of

reliable data, it is essential to pay attention to the testing and validation stages. The algorithmic bias that can arise during development is also a risk due to erroneous and excluding interpretations being induced.

One of the survey informants referred to algorithmic bias as a "malicious risk," which refers to the possibility of predictive models replicating the social, class, ethnic, or gender biases present in the training data. Correcting these biases during testing is key to ensuring that AI systems do not aggravate health inequalities.

> It may be that the model for detecting a certain tumor, for example, can do very well in the white population, but in the black population, it may not be so good. This is an inequity. It's like veiled discrimination without you realizing it because the historical data on training reflects this. There's this very strong discussion of historical discrimination or population representativeness within databases for AI training, and this, for example, is a potential risk. You must have the mechanisms to be able to check this and be actively looking at it. Does that mean that we have to stop, and you can't develop AI anymore? No, of course not. In fact, you have to keep doing it, but you have to have a whole team and mature science data to be able to mitigate and reduce these risks. Then we go back to that same old story: How do you mitigate this risk? If I have data that represents everyone. If the SUS, for example, is not a strong source of data for AI training, if only the private sector organizes its data, you're going to see that the tools will reflect the distribution and behavior of diseases in the wealthy population. That's what's going to happen.
>
> *(HEALTHCARE FACILITY STAKEHOLDER)*

The interviewees pointed out that if the data used to train the AI reflects pre-existing inequalities in the healthcare system, AI tools may perpetuate these disparities. For example, if certain population groups have outdated care histories, then AI technologies are unlikely to help with diagnoses and prognoses about them because, due to the lack of data on these groups, they cannot be trained on their characteristics. They may function as an instrument that worsens previously existing situations of neglect. On the other hand, if AI is trained with data that highlights certain benefited groups, it might improve services for these populations, thereby intensifying the inequalities between different segments of society and also accentuating Brazil's territorial and regional inequalities. While part of the population may experience substantial benefits, other groups may see existing gaps in their healthcare accentuated. The following quotes discuss the risks of reproducing and/or increasing inequalities and inequities:

We may make access inequality worse. There's a difference, for example, between the public and private sectors, with — let's say — personalized or predictive medicine being used much more in the private sector than in the public sector, where a lot more information could be collected. How are these models really being trained when they use a population that's not representative of the population of Brazil? There's a class-based bias in these private healthcare datasets.

(PUBLIC SECTOR STAKEHOLDER)

If the data isn't really representative of where it's going to be applied, and when I say it's going to be applied, I'm not suggesting it has to work for everything. I joke that these models are like the leaflet you get telling you how to use medicine, and we say: "Look, this is indicated for this purpose, and this is indicated for that other purpose. If it's contraindicated for you, don't use it." So, we need to know what the indications and contraindications are for each of these systems because they're definitely not for everyone, even though they're often talked about as if they were. So, there are very high risks of it being applied where it shouldn't because it's not going to work, and the system won't respond. And then we increase the inequalities that already exist in this country.

(ACADEMIA STAKEHOLDER)

[It's crucial] to ensure that models are trained with data from more remote regions of Brazil, where data collection is less frequent and fewer examinations are carried out. So, having this data quality, because if you don't have it, you risk training your algorithm using data from hospitals in wealthy regions that collect the most data. The algorithm might only learn to help with diagnosis and prognosis for patients in those regions. So, when you go to areas with different characteristics, it may not perform well. So, systematic data collection is essential, especially in regions where it's most needed, which are often the more remote areas of Brazil.

(ACADEMIA STAKEHOLDER)

Explainability of the models

The interviewees also highlighted the lack of mandatory protocols for testing and validating the tools as a risk. Some stakeholders argued that the algorithm development process needs to be transparent, undergo a mandatory set of tests, and be validated by a neutral body to receive a certificate or a badge. Otherwise, as one interviewee said, these tools might work like "medicine without its explanatory leaflet."

AI systems need to be thoroughly tested and validated. You can't just build a system, declare it ready, and expect everyone to trust it. It needs to be tested, validated, and subjected to various uses in order to build confidence. Otherwise, there's a risk of relying on incorrect information. I think the credibility of any AI project comes from extensive testing and validation.

(PUBLIC SECTOR STAKEHOLDER)

In response to your question about risk, I see the lack of a target or benchmark to measure my successes and failures as a significant risk when using AI tools. Let's imagine that today when I don't have AI, I adopt a particular approach with a patient, and it's proved to be the right approach. How do I know it's right? Because I've studied the outcomes. If I refer a patient down Path A, there's a 90% chance of improvement, whereas Path B offers only an 80% or 50% chance. Now I'm going to go back to using AI, and I'm going to ask you who's measuring its effectiveness? How can I be sure that by using AI like that, I really had a 90% success rate? [...] What was right was the improvement in the patient, in the patient's outcome; it wasn't the algorithm that was right. Not having a benchmark or an outcome measurement model is a risk, in my opinion. When a drug is introduced in the market, it undergoes a series of phases and rigorous testing before being marketed. In AI, the guy puts it on his computer or in his system, and that's it; there's no way of evaluating it. It doesn't get tracked and traced, and nobody follows any of the models.

(MARKET STAKEHOLDER)

The explainability gaps of AI tools were also emphasized as a potential risk, albeit by a limited number of respondents. Those who addressed the issue believe that the development and implementation of AI applications in healthcare are not accompanied by efforts to ensure transparency and the dissemination of the decisions and basic knowledge underpinning these tools' construction. They believe that if users do not understand the processes and decisions behind the construction of the algorithms, they will be unable to evaluate and criticize the decisions based on these technologies.

I think data governance is a risk. It's something that can become a risk for ethics. How do you explain the technologies that are being developed today? In other words, how do you explain the following: "This algorithm has been trained." But who trained it? On what basis? I see it like this – and in ethics, there's even a metaphor for it. It's as if someone had some medicine, and although there isn't an explanatory leaflet in the package, it cures their headache. Now, where was it tested, by whom, and what papers show this? You have that in pharmaceuticals, but you don't have that in AI today. We've been trained by so many patients in such-and-such places, by such-and-such universities. Here's all the traceability of who tested it, the name, the researcher, and everything. It doesn't have all that. I think it's very, very important to do this.

(ACADEMIA STAKEHOLDER)

Showing why the AI came to a certain conclusion helps a lot. Everyone who's ever been to school... you go to Math class, and you have that huge problem, and the teacher never accepts "Here's the answer." No, no, show me how you arrived at the result. So, we're at the point where everyone thinks that AI is going to always give you the best answer, like a calculator, but in reality, it's going to have to explain how it got there so that when you make a mistake, you understand why. [...] You need to show the flow of thought, and once the world understands that AI doesn't just give you the result, it gives you the flow, you mitigate a lot of risks because you're monitoring the rationale of the AI.

(MARKET STAKEHOLDER)

Automated decision making

Another risk pointed out by the interviewees is the inappropriate use of the information generated by AI decision support systems. They point out that it is essential to ensure that users understand that this data does not dispense with the assessment by healthcare professionals. The benefit of AI in this area is not that it replaces humans but that it helps with decision-making. Most systems are likely to generate answers that need to be evaluated by the healthcare professional before a diagnosis or referral can be made. So, professionals must be prepared to take advantage of the tools' benefits without neglecting their limitations.

> The greatest risk lies in the transition to effective application as a tool. As I see it, we're not mature enough in the healthcare environment today to guarantee that this is going to be properly implemented and its use monitored. So, the risk I see is more of misuse, in the sense that you're not guaranteeing quality control. We need to create a mechanism for understanding that this type of tool isn't something you just implement and leave. It has to be part of a context in which we understand that tools and technology in healthcare have to be monitored. [It requires] quality control and maintenance.
>
> *(ACADEMIA STAKEHOLDER)*

Accountability

Finally, ethics were also mentioned, albeit infrequently, in relation to questions about the risks of using AI in healthcare. When it appears in the narratives about risks, however, it is related to the individual, that is, to the ethics of the healthcare professional using an AI tool. The ethical guidelines for using this technology are not seen as collective constructions resulting from society's norms and discussions but as principles of individual ethics.

The discussion on ethics and risks provokes the debate about accountability because, at the current stage of AI development in healthcare, it is unclear which actor or organization is responsible for each part of developing and applying AI tools. This leaves unanswered a question posed by one of the study's interviewees: "Who's going to be responsible for any mistakes made?" Society and the medical community must address this complex question assertively. However, as the interviews suggest, it is still a discussion that has not gone into great depth.

To tell you the truth, I think it depends a lot on individual ethics. [...] I think that individual decision-making, or decision-making in public health that's based on findings and evidence produced by machines – I've got nothing against machines, they serve a purpose for us – but they need to be supervised.

(ACADEMIA STAKEHOLDER)

The ethical risk. If AI makes the wrong decision, who's going to be responsible? Those who made the AI? A doctor always has to sign, or a health professional. This has to be discussed before this ethical dilemma arises. Then there's the ethics of even those who are training this AI. How are you training this AI? If you train this AI with results, for example, you're training AI in lung X-rays. If the reports aren't reliable, it's going to give answers that aren't reliable. So, how do we assess this beforehand? There has to be regulation. Who are you training with? Who's doing these reports? How reliable are these reports? Who's going to be responsible for any mistakes?

(HEALTHCARE FACILITY STAKEHOLDER)

BOX HIGHLIGHTS - RISKS OF USING AI IN HEALTHCARE

- Consensus among stakeholders: The benefits of AI in healthcare outweigh the possible risks, but essential concerns still need to be addressed to ensure responsible and effective use of this technology.
- Regulation around the use of AI in healthcare: This is considered an area of concern by the different profiles but with many different views. Some interviewees believe it is necessary to create new regulatory mechanisms to guarantee the safety and reliability of AI tools in the healthcare sector. They also believe that broadening the scope of regulatory mechanisms could increase the costs and number of legal requirements associated with developing and applying this technology.
- Data privacy and security: Concerns about protecting sensitive data and the risk of information leaks are highlighted. Even with the LGPD, it is believed that there are regulatory loopholes that could compromise data security, mainly when new data is generated using AI tools.
- Algorithmic bias: There is a perception that there is a risk of AI algorithms reproducing and amplifying pre-existing inequalities in the healthcare system. If the data used to train AI reflects inequalities, the tool could perpetuate these disparities and have a negative impact on specific population groups.
- Transparency and explainability: Few interviewees explicitly mentioned that the results generated by AI systems should be explainable. However, many have pointed out the lack of transparency in AI algorithm development processes is a risk. Understanding the processes and decisions behind AI tools was also mentioned as essential for evaluating and critiquing their decisions.
- Inappropriate use of AI tools: There is a risk of healthcare professionals not fully understanding the

limitations of AI tools and relying excessively on their recommendations. Professionals need to be appropriately trained to use these tools to aid decision-making, not as a substitute for it.

- Ethical responsibility: The lack of an ethical framework with clear accountability and guidelines on how to act is seen as a risk for using AI in healthcare. Despite mentioning this risk, the interviewees tended to refer to AI ethics as the result of individual practices rather than collective constructions.

PRIORITY TOPICS FOR THE AI AGENDA IN HEALTHCARE

In the previous subsections, we provided an initial overview of the state of AI in the Brazilian healthcare sector. To complete this picture, we will now systematize the key themes from our interviews. These "key themes" are issues prominent in academic and government debates[7] on AI in healthcare. They also align with the DHS (MS, 2020) and are central to the AI agenda. These topics include interoperability, infrastructure, human resources, regulation, ethics, and user rights.

Given the relevance of these issues, it is essential to detail the methodological procedures used to gauge the interviewees' perceptions. As noted in this publication's chapter "Methodological Notes", the interview script was divided into two blocks. The first block included questions about the Brazilian context for developing AI in healthcare in order to gather perceptions about the current stage, opportunities, potentialities, challenges, and risks of using AI tools. In these initial prompts of the interviews, various spontaneous comments emerged about the six key DHS topics in most of the interviews. With the second set of questions, the script explicitly explored issues related to DHS guidelines.

To optimize interview time and prioritize the exploration of topics not covered by the interviewees, the following collection strategy was used: Comments on the issues of

7 For a discussion on the most discussed topics in the literature, see the article "Artificial Intelligence in healthcare: A view of the literature and guidelines for Brazil", by Rodrigo Brandão, in Part 1 - Articles of this publication.

infrastructure, human resources, and users' rights were encouraged only when these subjects had not been mentioned spontaneously in the interview up until that point, while the topics of interoperability, regulation, and ethics were encouraged with the aim of going into greater depth, even when they had already been mentioned throughout the interview. New reflections were often formulated based on this dynamic, so the topics covered in this section may revisit some of the issues from previous sections. The aim is to briefly systematize stakeholders' perceptions of these critical issues for the academic debate on AI, healthcare, and DHS.

BOX 3 - DHS

The objective of the DHS is to organize and strengthen actions in digital healthcare. Its purpose is to guide public and private initiatives in driving digital transformation in Brazilian healthcare. Three main lines of action have been outlined for achieving the Digital Health Strategy Action Plan: (a) Actions by the Ministry of Health for the SUS, in particular the Conecte SUS program as a crucial part of the digital healthcare vision; (b) defining guidelines for collaboration and innovation in digital healthcare, with an emphasis on expanding and consolidating governance and the necessary organizational resources; and (c) establishing the DHS Collaboration Space in the search for an efficient exchange between all the players in the sector, who have defined roles and responsibilities.

The plan also has seven priorities: (a) governance and leadership in digital healthcare; (b) computerization of the three levels of healthcare; (c) support for improvements in healthcare; (d) user empowerment in digital healthcare; (e) training human resources for the area; (f) establishing an interconnected environment; and (g) developing an innovation ecosystem in digital health (MS, 2020).

Interoperability

Interoperability is a central issue on the agenda for using AI in healthcare. The theme emerged very strongly and spontaneously in most of the interviews. Stakeholders mentioned the interoperability issue in Brazil as both an opportunity and a challenge. Regarding specific opportunities for Brazil, the fact that the country has a single public health system with a large volume of data is particularly important.

Promoting interoperability between data in this system is considered a unique opportunity because of the volume and diversity of this data. The interviewees believe that

integrating the SUS databases could lead to the development of robust algorithms and boost the development of this technology in the country.

The RNDS is perceived as a positive strategy for promoting the integration of healthcare data and systems in the future. There is an understanding that the network can operate not only as an important articulator for integrating data in the public sector but also for including data from the private sector since the competition and the lack of general guidelines make integrating information in this sector difficult.

Despite this potential, the lack of quality data and the difficulty in establishing interoperability are among the main shortcomings identified by stakeholders. They point to the lack of connection between systems, the difficulty in developing protocols to standardize data, and unequal resources for registering the data in the different territories in Brazil. The interviews emphasize that the lack of uniformity and standardization in health data compromises interoperability, making it challenging to train algorithms effectively. These shortcomings bring risks as they increase the potential for algorithmic bias. Limited databases or those with a concentration of information relating to specific population profiles increase the chances of biased analysis, thereby weakening the quality of what is produced by AI and, ultimately, potentially jeopardizing access to healthcare and the quality of clinical diagnoses and referrals.

Infrastructure

Adequate infrastructure for the development and application of AI is unevenly distributed across the country, with a more significant presence in large urban centers. The interviews highlight that many municipalities face precarious situations, with a lack of specialized laboratories for developing and testing AI applications in healthcare and a shortage of equipment for digitalizing data in healthcare facilities. Therefore, this precarious situation is a barrier to advancing the use of AI in healthcare in the most diverse territories and local contexts.

Human Resources

The lack of human resources trained to implement and use AI in healthcare is a significant challenge in Brazil. The

interviews reveal that it is tough to recruit professionals with a minimal understanding of AI and healthcare, such as data scientists who work with health data or healthcare professionals with a background in technology and programming. They also point out that the increase in demand for IT professionals has led to an increase in the salaries of these professionals in other sectors (such as finance), making it challenging to retain specialized professionals in the country.

Regulation

The interviewees' discussion about regulating AI applied to healthcare is complex. On the one hand, there is a demand for greater legal clarity on what can and cannot be done about diverse topics involving AI in healthcare. On the other hand, the creation of rules for AI systems is viewed with trepidation since, according to various interviewees, they could discourage the progress of these systems in the country, leading to Brazil "falling behind" in the race to develop and implement AI in healthcare.

Although the LGPD was not enacted with this intention in mind, according to the study's interviewees, it is an essential instrument for successfully developing and applying AI tools in Brazil. It is considered an initial regulatory framework and the basis on which applications can be developed, tested, and applied. However, as it did not consider the particularities of AI technology when it was drafted, gaps in its regulatory framework leave it open to different interpretations of its practical applications and legal limits. Although the LGPD regulates the use of personal data, it does not apply to any new data created from existing personal data or to what happens to them after people die. A few interviewees pointed out that this could pose a risk to data protection and security when implementing the technology. Therefore, the absence of specific regulations for the development and application of AI in healthcare can have ethical, legal, and data security implications.

Ethics

The discussion on regulating the uses of AI is associated with the debate on ethics and AI, but, as mentioned above, it did not appear spontaneously and strongly in all the interviews: With a few exceptions, the question of ethics only came up when interviewees were explicitly encouraged to

consider it. All the interviewees were asked about the ethical implications of using this technology. Still, they generally did not reflect on the issue in depth, and many did not clarify what they understood by "ethics" in the AI debate. As a rule, this discussion is linked to the values of the individuals who develop or operate AI tools, i.e., the ethical framework for dealing with the implications of using AI in healthcare depends on the individual ethics of the professionals involved. When they can envision collective or social mechanisms for the ethical ordering of the use of these technologies, they talk about strengthening the topic of ethics in professional training courses and existing professional codes of conduct.

Since AI would be just a working tool like any other available tool, general healthcare ethics training, as set out in the professional codes for doctors, nurses, and other healthcare professionals, would be sufficient. Many stakeholders believe that working with AI is similar to any type of empirical or clinical study and should follow the same ethical precepts as research and medicine, without specific ethical guidelines; ensuring data privacy and applying medical principles would be sufficient. Validation of the ethical procedures for using AI in healthcare could also follow the same dynamic as the validation of empirical studies, generally based on debates in forums involving healthcare professionals and assessments by academic referees.

In fact, I don't think we need to invest to work with ethics and regulation in AI, like radically changing the way it's done, creating something. No, the rules are the same as they always were, but they have to be clear. So, there are various manuals out there on good conduct, on how to do research and adjust the bias in AI, only doing it with a team that's trained to do this research checklist. I think it's a matter of education; it's really education. We have to do a lot of work on this with researchers and in the industry.

(HEALTHCARE FACILITY STAKEHOLDER)

If you handle patient data and use it with a tool, it should be treated in the same way as the code of conduct that doctors adhere to when they become doctors, that nurses adhere to when they become nurses, or that hospitals adhere to when they start operating. It should be the same; that should be enough because it's the same thing — it's just the tool you're using. When a lawyer is governed by the Bar Association (OAB), no one says, "Oh, I want to know the code of conduct for Dell or Lenovo, the companies that made the computer you used to write that law or defense." No one does that; it's just the tool they're using. The Bar Association already exists to oversee the final use and outcome, not the tool itself. 'Oh, but the tool needs to be looked after.' Of course, it has to meet all the security criteria there are, just like any laptop has to, and each market has its own standards. You don't need to create something massive.

(MARKET STAKEHOLDER)

Ethics is a question of training. So, we need to work on training professionals in the healthcare sector, the exact sciences, engineering, those who work in this environment. [...] I believe this question of ethics needs to work with both sides [IT and health]. It's professional training for those who work in this area.

(ACADEMIA STAKEHOLDER)

The first point is to go through the competent forums to decide on this; the second point in this process is that you have to guarantee data anonymization to ensure approval for a research project in this area, the *ad hoc* [reviewer] will certainly look at the issue of equitable distribution among representatives of the population.

(ACADEMIA STAKEHOLDER)

It's the people who are responsible for any ethical considerations, not the system itself. The ethical aspects come from the individuals developing the system, who must question whether their work infringes any ethical principles during the development and implementation phase of the model. So, I believe that ethical aspects should be inherent in the people involved, rather than expecting them to be embedded in the model itself.

(ACADEMIA STAKEHOLDER)

Among stakeholders working in academia and healthcare facilities, the discussion about the ethical implications also includes a concern about the principles that guide the development of AI tools. They stressed that the use of AI will be ethical if it is equitable, that is, if AI algorithms and systems are developed fairly and impartially, and if their results do not contribute toward increasing previously existing social inequalities. The importance of ethical concerns in all algorithm development processes was also mentioned, and whether or not the criteria used by the models are in breach of ethics should also be identified.

Addressing ethics involves implementing an equity scheme, giving more to those who need it most, and then developing an algorithm that will help ten people. I think we have to care for each other and give more to those who need it most and do good rather than cause more harm.

(HEALTHCARE FACILITY STAKEHOLDER)

User rights

Finally, it is essential to note that although the importance of guaranteeing the confidentiality and security of patient data was mentioned by most of the interviewees, the discussion about the rights of the ultimate beneficiaries of technological tools in healthcare (the patients) did not appear spontaneously in most of the interviews. Although the rights of the users and beneficiaries of the tools are a central theme in DHS, it was only superficially mentioned by the few interviewees who

touched on the subject. When they were encouraged to talk about this topic, their discourse was short, lacked depth, and often returned to data security.

A significant number of those interviewed believe that the rights of health system users are limited to data confidentiality and protection. They talk about the importance of people knowing more about the provision of personal data in healthcare in order to reduce fears and resistance, and to understand better how the LGPD is used. They also argued about the importance of having digital education strategies for the population so people can understand the benefits of technology and the potential uses of the data they provide. When they were asked about the rights of the users of healthcare systems, one stakeholder summarized their perception as "the right to have the benefits of using AI," while another replied that it was "the duty of citizens to provide their data so they can benefit from it." Over and above the right to data confidentiality and protection, the interviews did not explore other user rights.

It's something that's going to benefit him at some point in his life, isn't it? So, I think users have a duty rather than a right to provide their data. I believe they have to share it; they need data to make the algorithm work well. What you can't do is leak data... guarantee the anonymity of the person, but I can guarantee that they'll never ever be recognized. It'll be impossible to reach that person, and this will only be used to do something beneficial to the healthcare network. I believe it's more a public duty. You just guarantee anonymity.

(HEALTHCARE FACILITY STAKEHOLDER)

I'm of the opinion that if it's collectively and individually beneficial, then we shouldn't impose excessive limits. I don't think we can. Our individual limits shouldn't outweigh the collective benefit.

(HEALTHCARE FACILITY STAKEHOLDER)

Finally, we note that the issue of explainability of AI tools, which is so relevant in the academic discussion on AI and healthcare, was also not a significant concern for most of the interviewees in the study. The lack of transparency and understanding of AI applications' decision-making and implementation processes concerns a small proportion of those interviewed. This indicates that some of the priority topics in this field of study, such as explainability, reliability, and user rights, have not resonated with stakeholders in this field in Brazil.

BOX HIGHLIGHTS - PRIORITY ISSUES FOR THE AI AGENDA IN HEALTHCARE

- Interoperability in Brazilian healthcare is considered an opportunity due to the large volume of SUS data, and the potential of the RNDS to promote the integration of public and private data.
- At the same time, interoperability is considered a challenge due to the difficulty in putting what is necessary for it to happen into practice. Uniformity, standardization, and equal registration in the country's different territories are essential if the results of AI technologies are to be reliable.
- Economic and social inequalities create gaps in infrastructure and human resources for AI in healthcare. There is a lack of resources in municipalities and local contexts, such as laboratories and facilities for digitalizing data, compared to what happens in large cities across the country.
- There is a significant challenge in recruiting trained professionals to implement and use AI in healthcare in Brazil, such as data scientists and healthcare professionals with expertise in technology.
- The regulation of AI in healthcare is complex. There is a demand for guidelines on data processing, but there is also a fear that the measures that will be put in place will restrict and, therefore, slow down the country in the technological race. As much as the LGPD is considered a relevant initial regulatory framework, its lack of consideration for the particularities of AI raises ethical, legal, and data security concerns.
- The discussion of ethics in the regulation of AI in healthcare only came up when explicitly encouraged in the interviews. In these cases, the most recurrent view is that ethics depend on the individual values of professionals, suggesting that professional training and existing codes of conduct would be sufficient. In addition, most interviewees did not clarify what they understood by "ethics" in AI debates. The few who did so emphasized the importance of fairness when AI tools are being developed.
- The topic of "users' rights" was not central to most interviewees; the only highlights were data confidentiality and protection and the population's lack of digital education to understand its uses. The question of the explainability of AI tools was also not widely discussed by the interviewees.

PRACTICES IN PROGRESS

The interviews also sought to identify examples of the use of AI systems in healthcare in Brazil and actions aimed at developing this technology in the country.

Initially, the interviewees' comments on the current stage of development and application of AI initiatives in their organizations are described and analyzed. These highlight the most commonly adopted initial strategies considering the most frequently perceived initial barriers.

The initiatives underway will then be described in terms of their objectives. We mainly mapped those initiatives that had four purposes: (a) promoting interoperability, (b) solutions for improving management, (c) diagnostic tools, and (d) prediction models.

The section closes by highlighting the challenges and risks that were pointed out by the different stakeholders in developing and/or implementing the initiatives mentioned.

IMPLEMENTATION SCENARIO AND THE CURRENT STAGE OF INITIATIVES IN BRAZIL

Analysis of the interviews reveals that existing AI initiatives in the different segments investigated (academia, public sector, the market, and healthcare facilities) are at an early stage of development or application and face similar structural challenges.

Regardless of the segment in which the interviewees operate, the reports revealed that existing AI initiatives were initially based on the institutions' own structures, such as laboratories, intelligence centers, and specific departments. Organizations are concerned with developing minimum structures and an implementation environment geared explicitly toward AI initiatives in healthcare.

The study's interviewees also emphasized the importance of forming partnerships with players from other segments to develop AI solutions and foster an innovation ecosystem. The most frequently mentioned partnerships were between healthcare facilities, academia, and technology companies. Partnerships were mentioned less often in the case of the public sector and when they were mentioned, they generally dealt with relationships between different levels of government or with universities and, possibly, startups.

There's a huge variety. There's the possibility of startups and companies teaming up with university laboratories. So, there are companies in the bioinformatics area that work with genomic medicine and have partnerships with the genetics department of laboratories, for example.

(ACADEMIA STAKEHOLDER)

The interviewees highlighted the need for financial and human resources as a critical factor in encouraging partnerships seeking funding. Looking to funding agencies for financial resources, for example, is a well-established practice, not only in academia but also among health equipment companies linked to universities and among startups.

We're developing technology in the healthcare area for various companies. These range from those making medical equipment and that produce this technology in the country to our own laboratory that develops software to improve the efficiency of this equipment. So, most of the lab's funding is from projects that come under IT law, but we also have a regular line of research with [funding agencies], specifically in AI.

(ACADEMIA STAKEHOLDER)

On the financial side, which is a very important backup in terms of investment [...] is the [Development Agency], which has an incredible program, [...] [they provide] the funding that the company needs to get an idea off the ground or validate an idea with the market and build a prototype, a solution.

(MARKET STAKEHOLDER)

Concerning the specifics of their work, it was impossible to distinguish specific types of AI solutions at a more advanced stage of development in any of the segments we interviewed. We were able, however, to identify different stages of implementation connected to the interviewees' areas of activity. Many initiatives have been implemented or are at more advanced stages of testing with healthcare facility and market players. From the interviews with these segments, it was also possible to map the use of AI solutions implemented in other countries and included in the daily work of some healthcare facilities in Brazil.

PROMOTING INTEROPERABILITY

In all the interviews, interoperability appears to be a central theme because it is a critical element in the development of AI. Greater integration between different data sources leads to better-quality solutions. As mentioned in the previous section, although it is a priority process for all sectors and necessary

for the advancement of AI in the country, its development is quite complex, as reported in the interviews, as it requires the articulation of public and private players, investment in infrastructure and, in some cases, the transfer of data between competitors, which can generate conflicts of interest.

The interviewees mentioned different efforts and measures to ensure interoperability. In terms of initiatives, the efforts of government bodies to digitalize and integrate health systems through the RNDS were highlighted. However, challenges related to the adoption of national standards and the engagement of all stakeholders still exist. In addition to these efforts, the interoperability initiatives led by the various segments interviewed have different maturity levels, resources, and limitations to their implementation.

The federal government has outlined institutional changes that align with the guidelines of the DHS and the National Information and Informatics Policy and contribute to interoperability. Among the changes is the reformulation of an MS department to monitor and evaluate the digitalization and use of AI in Brazilian healthcare. Councils and bodies connected to this ministry have also dedicated efforts to developing information exchange and institutional networks between government entities to establish a basis for dialogue about advancing the use of AI and innovation in the SUS.

> There's the history of the DHS. This document, which is a fundamental milestone, is linked to the National Information and Informatics Policy. [...] It's a fantastic opportunity. Institutional changes are occurring, including the restructuring of the [evaluation department] and the new secretariat, which is also an important milestone in the country. We're revisiting old elements but with a fresh perspective, such as the Strategic Management Support Room and the Interagency Health Information Network. In other words, more than anything, these are bridges to dialogue.
>
> *(PUBLIC SECTOR STAKEHOLDER)*

The MS has also invested in interoperability by preparing and developing mechanisms for integrating data from different public patient information systems and databases. This process is behind the *Conecte* SUS platform,[8] which provides patients, professionals, and managers with information about patients and their care.

8 Read more: https://www.gov.br/saude/pt-br/composicao/seidigi/conecte-sus/conecte-sus

We're adopting an interoperability architecture model in the Ministry of Health, known as the Fast Healthcare Interoperability Resources (FHIR) standard,[9] which is the national health data network. We're gradually working on processing, downloading, and enriching the data. We're now going to start trying to return the processed data to the states and municipalities so they can also use it on different levels of aggregation. This allows us to have and produce Conecte SUS professionals and managers for these citizens: The citizen's Conecte SUS, and the professional Conecte SUS, which deals with electronic medical records, and the Conecte SUS manager, which is information aggregated in the form of indicators for monitoring and evaluating public policies. Brazil has taken a major step with this RNDS interoperability model, and we're involved in this process. We're now in the process of downloading the regulation, which already has an information and computation model and the minimum established data. The medical records already have vaccination data, laboratories..... all of this is already being downloaded into the RNDS.

(PUBLIC SECTOR STAKEHOLDER)

Even though it is still being developed and facing the challenges of implementation in a country of continental dimensions and marked by various access inequalities, the federal government's efforts are an essential step toward building higher-quality information repositories with fewer risks of bias against the population. However, because, at present, there is no consolidated repository with regulated access or joint construction by different stakeholders, interviewees mention efforts to connect patient data as necessary for developing their AI projects. Interviewees from all sectors, including the government, mention different strategies for developing repositories based on the data they have access to.

Academia, the market, municipal governments, and healthcare facility managers reported using different repository formats and ways of operationalizing the integration between different data sources. Interviewees from municipal governments, healthcare facilities, and the market cited the use of data lakes, i.e., a central repository that combines structured, semi-structured, and unstructured data, with or without identifying keys, as an instrument that can be used to develop data interoperability.

We started Digital Health in 2021 with this in mind: We need an electronic health record to integrate these databases. We've managed to implement [...] a data lake. We have a national health record of the entire health system of the municipality that makes connections, including with the private sector, It connects other networks.

(PUBLIC SECTOR STAKEHOLDER)

9 A standard for exchanging health information.

Another thing we've been doing for several years now is a strategic project for organizing and structuring our data. For example, today we have a data lake with lots of healthcare and laboratory data on the Brazilian population. Our group is nationwide, so it has had a unified data system for many years now.

(HEALTHCARE FACILITY STAKEHOLDER)

Over and above the repository format, interviewees reported different operational approaches for integrating data. For example, data buses have been implemented by public sector actors and healthcare facilities. This involves investing in infrastructure and developing keys that enable the interconnection between the data used in different software systems. This procedure is used by various organizations that were mentioned by the interviewees to integrate and centralize information into a single database.

When we work with the DHS in [the municipality], our database needs to connect this information. We're producing and collecting a lot, but if we don't organize it according to an interoperability standard with a data bus and with the National Health Network, it's also going to reflect on our ability to integrate the databases that are here into information models. These were not produced by the RNDS; this is the starting point.

(PUBLIC SECTOR STAKEHOLDER)

We're currently in the process of creating our data bus. We have a main repository (MR) and around 50 software systems that surround this MR but don't communicate with each other. Our first major step was to create this data bus to enable interoperability among all these software systems so we could centralize the data into a single database.

(HEALTHCARE FACILITY STAKEHOLDER)

Interviewees from academia also talked about building database integration systems, including those that rely on a common identifier and cases that develop algorithms that connect different databases without unique identifiers and using encrypted, non-reversible codes to guarantee the anonymization, confidentiality, and protection of personal data.

> We've been using a tool, which is a probabilistic algorithm designed to connect databases that have no unique identifiers. It uses data such as the name, mother's name, and date of birth, but as follows: It encrypts the data by transforming names into nucleotide DNA triplets. Each letter is then converted into a DNA triplet as if it were a sequence. Each time a patient is entered, they are encoded with a different, non-reversible set of nucleotides — you can't reverse the conversion. So, they [the algorithms] encrypt sensitive data, which, in theory, resolves part of the anonymization problem. With these sequences, we can compare data with those of another database using the Basic Local Alignment Search Tool (BLAST) algorithm. This addresses the issue of interoperability and connects systems and databases that are not otherwise linked, allowing us to achieve a minimal level of interoperability. We've been trying to offer this to the Ministry of Health and secretariats for them to use at no cost to them.
>
> *(ACADEMIA STAKEHOLDER)*

MANAGEMENT TOOLS

The use of AI to develop management tools is wide-ranging. Interviewees from the market, healthcare facilities, and academia cited investment in this type of initiative.[10] The interviews showed, however, that management initiatives are an investment area, especially for startups and better-established technology companies. Unlike diagnostic tools, we identified solutions among management tools that are being implemented daily in hospitals, laboratories, and other healthcare facilities.

The ongoing initiatives mentioned by the interviewees involve patient care, the optimization of internal operational processes, and different devices that sort and organize queues. Regarding care, there has been investment in tools that use generative AI in chatbots to improve patient adherence to care and monitor the post-operative period. An interviewee from a healthcare facility has developed an AI model that guides bone marrow transplant patients through the postoperative period. It indicates what they should do, what kind of restrictions they must implement in their routine, and when they should contact their doctor.

AI has also been used to record or generate data for those providing the service. The interviewees mentioned initiatives that retrieve details of the user's journey and, in an automated way, give a history of their hospitalization, for example, to the different hospital staff who need to be involved. There are also

10 Among those interviewed from the public sector, no initiatives were identified that contribute to health management. There was, however, mention of the use of data categorization tools to search for information on bills.

at least two initiatives for recording information: One uses the patient's written online care details to build the medical record, thus supporting the healthcare professional in their decision to refer the patient for face-to-face care or telemedicine; the other uses the unorganized oral report of the analysis of an image to record information in reports. The former has been used in a health plan and, due to machine learning techniques, has had good registration results.

> We have [...] what was active was the medical record system, so we have a summary and make a consultation note. What should we call this? I do a pre-registration – that [sounds] better – I do a nursing pre-registration. [...] [With that], I reduce the time spent on the medical record system, but it hasn't happened yet, and I also get a very strong "quali," with a very good user experience. This was the first prototype we managed to build from the medical record perspective. [...] [The pre-registration] by all the nurses, for all the appointments, 500 appointments a day using AI for pre-registration in the "Symptoms (Subjective)" and "Objective" fields. We use a methodology called SOAP [, which requires recording Subjective-Symptoms, Objective, Assessment, and Plan], where symptoms and objectives are auto-filled, but the "Assessment and Plan" sections aren't yet. [...] We handle this using chat, so it continues to operate. When it goes to register, it says: "auto-register." It takes the conversation the nurse had with you in the chat, uses the prompt we've constructed, structures it technically according to the SOAP methodology, and presents it as a suggested record. The nurse then adjusts the text and publishes it, or generates it again, or starts from scratch.
>
> *(MARKET STAKEHOLDER)*

> We have a project that aims to improve the efficiency of radiology reports so that the radiologist can dictate the report instead of typing it. He dictates the information, and AI transforms it into a standardized [report]. This avoids information being different from the radiologist to the radiologist, and then a radiologist must review the report and sign it. This would help standardize radiology reports.
>
> *(HEALTHCARE FACILITY STAKEHOLDER)*

Some startups are also developing systems that use AI to assist in remote care, such as tools for analyzing symptoms and making preliminary diagnoses based on the patient's report. One interviewee reported that the startup he works for has built an AI tool for triage aimed at remote consultations. It tries to identify the patient's requirements and performs a triage for the type of care needed.

In addition to the applications being tested for triage, algorithms are also being developed to organize care queues and optimize internal operational processes. One university is developing programs in partnership with companies for administrative issues, such as queue management, and developing Robotic Process Automation (RPA), i.e., robots for operational tasks.

One corporate initiative uses AI to map the patient's journey and communicate the problems identified to the professionals involved, connecting information between different sectors and seeking to automate standard authorization procedures.

> We also strongly believe in AI as part of the business itself, for queue management, administrative management, RPA... we've been testing these in some regions.
>
> *(ACADEMIA STAKEHOLDER)*

> Here at the company, we digitalize the patient's entire journey in the hospital. Anyone who goes into the hospital has already gone through check-in and the paperwork, whether it's for a consultation or they've been admitted for surgery. At each of these points, we have a module that digitalizes part of the operation. What operations do we digitalize? It depends on the point. It's always going to be that process that has the most bottlenecks. So, if I'm checking in, or at the documentation part, checking with the healthcare providers if it's for hospitalization... there's a big problem with hospitalization, which is the following. During the operation, in the operating room..... you're undergoing surgery and there's something missing right there and then, and a nurse has to leave the room and run to get some piece of equipment, or a bandage, something that's missing. Our company has a solution that it places in the room to facilitate communication with this nurse and the areas in the operating room. In hospitalization, as I said before, we see bottlenecks in terms of the efficiency of the nursing team, so we put our solution in place to facilitate communication between the patient and the areas.
>
> *(MARKET STAKEHOLDER)*

> My focus [...] is on facilitating access to authorization quickly through automation. [...] We have an organized database, so I already know how institutions behave, and I know the risk level of individuals, where they can spend more or less time in the hospital. [...] I'm currently using analytics to study this risk. So, I recommend that the audit team review it, because the data can't make decisions on its own. You need someone at the top to make the decision with the information they have, and most of the time, they get it right, but sometimes they don't. At the same time, I've developed the same kind of information for doctors, so both the doctor and their assistant have this information, too.
>
> *(MARKET STAKEHOLDER)*

Also, about management processes, one of the interviewees mentioned an AI system that is being developed to facilitate communication between teams and increase efficiency based on analysis of the customer's medical care journey. AI records the patient's entire journey, from arrival at a facility to clinical outcomes, including documentation, examinations, medication, and surgeries, and uses this information to coordinate the different sectors during care in a more automated way.

DIAGNOSTIC TOOLS

The fieldwork revealed that a significant part of the initiatives underway is dedicated to developing algorithms for diagnostic analysis. Above all, universities, technology

companies, startups, and healthcare facilities have invested in image-based diagnostic solutions. For example, a university was indicated as using AI tools for image analysis to formulate reports, prioritize care, and triage. These initiatives combine in-house development and solutions already implemented in other countries. One example is a company that has introduced automated image analysis for X-rays, chest CT scans, and prostate and skull magnetic resonance imaging (MRI) scans in its Brazilian branches. This initiative has been tested and is already being marketed in Brazil.

Automated image analysis initiatives cover various records, including tomography, radiography, MRI scans, and histology. Interviewees from different sectors highlighted, above all, the development of algorithms aimed at specific pathologies. One example is an initiative that is still being tested in a public university to develop an algorithm for diagnosing patients who have suffered a stroke. It can classify cases as either hemorrhagic or ischemic, which require different treatment routines and medication. Initiatives have also been reported in healthcare facilities for detecting liver tumors and molecular alterations (biomarkers) in images of human tissue and cells (histological).

We also have various initiatives being carried out here, such as the work that focuses on imaging. We work a lot with neurological imaging: Neurology is a very strong area here at the university and in the hospital. Then there's research into local things, systems, and local projects, but there's also partnership work with companies that are testing solutions in the hospital to assess patients who have had a stroke, for example; to see if it's hemorrhagic or ischemic so that they can be medicated. The emergency unit here has a stroke center that's a benchmark in the region. They work on applying methods to support diagnosis. This is also true in other areas, such as oncology and hematology.

(ACADEMIA STAKEHOLDER)

We have a very interesting project that was recently reported in the press, in which AI is being used to detect liver tumors in patients with chronic diseases, such as cirrhosis, a chronic liver disease. They are high-risk patients for detecting liver tumors, and we've developed an algorithm that helps the radiologist detect these tumors using imaging.

(HEALTHCARE FACILITY STAKEHOLDER)

We're very interested in detecting biomarkers directly from the histological image. So, cancer is a genetic disease driven by mutations that occur in the DNA of cells. More personalized patient management is based on detecting biomarkers, which are the measures that are evaluated. These biomarkers are detected through genetic test analysis. It's molecular biology, but these tools aren't always available — these technological tools in laboratories, at least not in most hospitals. Our idea is to use the histological image to detect whether the patient has the biomarker and whether they have the deficiency, for example.

(HEALTHCARE FACILITY STAKEHOLDER)

The development of AI related to specific diseases is partly the result of medical specialties with a solid history of research into healthcare facilities. However, it also reflects the degree of maturity and availability of databases, which are fundamental to the evolution of algorithms.

Among the initiatives that seek to create more comprehensive algorithms, we highlight the efforts of a university's laboratory to develop algorithms for analyzing and diagnosing images in radiology more broadly. This advance was made possible thanks to a database built during the pandemic, which brings together detailed patient information, including images and laboratory tests. This repository has been essential for developing AI solutions that produce diagnostic reports by identifying relevant image patterns and characteristics. These imaging algorithms have been tested, validated, and are being marketed. According to the person we interviewed, the results are of comparable quality to those of the leading companies on the market.

We set up a laboratory. [...] Here, we use a lot of [AI] in imaging, in radiology. So, we're already testing and developing the algorithm, and we even have a marketplace. But we're placing our bets on a few key areas, one of which is leveraging AI in imaging. By integrating AI, we aim to speed up reporting and boost efficiency in this field.

(ACADEMIA STAKEHOLDER)

Finally, diagnostic analysis has also been used for case prioritization and triage, which combine diagnostic tools and management solutions. One company described the development of tools that analyze images and patient histories to help screen and prioritize cases and identify and diagnose chronic or high-risk diseases early on.

Today, we have two product lines. One of them is image analysis with computer vision, where we screen patients in emergency rooms, in telecare, and pre-analyze the medical images. [...] We have another product where we view the patient as a whole by examining all their records. For example, if a patient has visited a doctor, a nutritionist, and an orthopedist, and mentioned that they're taking a medication meant for diabetics — although it has nothing to do with their knee pain — our algorithm can identify this. It might suggest, "Oh, is it worth scheduling this patient with an endocrinologist to manage their diabetes, check if everything is in order, ensure they're on the correct dosage, adjust his treatment, or perhaps should we just arrange a follow-up with a nursing assistant to see if they're taking the medication correctly, if they're well-trained in it, and if they know how to measure their glucose properly?" These are simple things, but they've become our main focus.

(MARKET STAKEHOLDER)

PREDICTION MODELS

As in the case of AI solutions for diagnostics, the interviewees reported using imported and self-developed prediction models. In the case of the former, a tool was mentioned that aims to predict which pathologies patients may develop and even generates an order in which the chances of them may occur. Although the algorithm was developed from other populations, it was implemented and is used today by a healthcare facility in Brazil that makes use of an integrated repository of Brazilian patient data.

Regarding self-development, the interviewees mentioned two event prediction models. One measures the possibility of an epidemic breaking out, while the other assesses the progress of a communicable disease and tries to identify its following territorial foci. These initiatives were developed by academia and a team from the public sector, respectively. In both cases, the interviewees stressed the importance of prediction in order to prepare for and possibly contain the event in question, including thinking about the operation logistics of the response and optimizing the allocation of resources.

We have a new initiative here to make use of electronic medical records produced by doctors and nurses in primary care, in the Emergency Care Unit (UPA), in the emergency network, for health surveillance, and for detecting outbreaks and epidemics quickly. Nobody uses paper medical records anymore, except in very specific places. It's a wealth of data that has the clinical description of the patient, and this data is ignored. Nobody uses it. This project aims to apply models to this text to classify it into a set of syndromes that have epidemic potential. The idea is to apply the models and classify and monitor them using three categories: person, time, and place. The idea is to use this data, classify it into syndromes to produce real-time surveillance information, and inform the healthcare manager that something is happening; not [that it happened] a month ago, but today, or yesterday at the latest, and that it's in neighborhood X, between streets Y and Z, and it appears to be dengue fever. Then, he knows what measures to take. It's a kind of epidemiological warning using clinical data.

(ACADEMIA STAKEHOLDER)

One of the things the [intelligence center connected to the health management area] did was create an epidemiological curve based on spatial optimization to predict where, in [the capital of a Brazilian state], there would be more cases over the next four weeks. This analysis identified which areas would likely experience a higher incidence and effectively tracked the movement or migration of cases. It was a significant effort in spatial optimization.

(PUBLIC SECTOR STAKEHOLDER)

Risk prediction initiatives that involve patients more directly were also mentioned. Three cases cited by the interviewees stand out: (a) A project by a healthcare facility that predicts breast cancer at an early stage, based on laboratory tests; (b) a project by a state government that predicts the risk of death in newborns, which is still at the pilot stage; (c) and a project under development by a company for monitoring and predicting patient falls. In all three cases, the interviewees stressed the importance of building and maintaining an extensive database in order to develop the algorithm; in the last two cases, the interviewees reported that the systems used showed positive results and helped with the prioritization and allocation of resources.

[...] that focuses a lot on laboratory imaging tests, so we have a lot of data today. I'll give you an example. We recently developed a dashboard with data from the breast cancer screening tests of women spanning over 20 years. So, we have data, for example, that shows the prevalence of breast cancer according to age in different states. [...] I've been leading a research project to develop a predictive model of breast cancer risk based on routine laboratory tests. [These are] common laboratory tests, for instance, such as blood counts, cholesterol levels, and lipid profiles, which are linked to breast cancer risk.

(HEALTHCARE FACILITY STAKEHOLDER)

We have a project that started at the end of 2018, which is only now taking shape; it's being implemented and even winning awards. It's a project aimed at reducing infant mortality. [...] So what does the project do? It captures various characteristics of a newborn; everything from gestational age and the number of prenatal visits, to birth details such as weight, any congenital malformations, and whether the baby was born prematurely or not. It goes to the database of live births and finds children with those characteristics. And then I take those children and go to the one-year mortality database. Did the children who were born with these conditions survive, or did they die within a year? How many died? The solution makes this prediction for a child who is born and receives this assessment. The nursing technicians can identify the issues: "This child might need 20% more attention." Along with this assessment, the solution includes a range of care recommendations, such as cleaning the umbilical cord, bathing the baby, checking the temperature several times a day, administering vaccines, and other specific care instructions tailored to suit each child's condition.

(PUBLIC SECTOR STAKEHOLDER)

The big solution we have today is inpatient monitoring systems for the risk of falling. So, we know that the risk of falling is an adverse event. It should never happen, and when it does, as well as impacting the patient, it also impacts the institution in terms of its view of the market; you can't gain any accreditation/certification from any foreign body if your level of falls is high. Today, our company has cameras on the beds and, using computer vision, we can detect the patient's position before they get out of bed. So, we can detect the patient's risk of falling by cross-referencing the data in the medical records. If I'm in my twenties or thirties, I haven't taken any medication in the last six hours, I've had a procedure on — I don't know — my arm, I can sit up in bed. But maybe John, who had a procedure on his knee, who's in the same age group and took x medication three hours ago, can't. So, that's already a risk of falling. And then the great thing is that today this tool is not only trained with real images, but it also generates its own images based on real images.

(MARKET STAKEHOLDER)

CHALLENGES IN IMPLEMENTING AI IN HEALTHCARE

We can identify some recurring themes when we look at the challenges faced by the different sectors involved in implementing these AI initiatives. Regarding financial resources, interviewees from academia and the market pointed to the difficulty of investing in AI projects. They report that there are few lines of public investment for autonomous development of this type of technology. In the case of academia, the difficulty was also reported in the fact that the existing lines of work are sometimes conditional on partnerships with private institutions, which could lead to significant changes in the work, such as the autonomy to carry it out and conflicts associated with the interests involved in developing solutions.

Interviewees from all segments mentioned challenges relating to human resources. Academics, healthcare facility managers, entrepreneurs, and public managers point to recruitment, specialization, and maintenance as recurring challenges when implementing their initiatives.

Having the funds to maintain the teams is seen as a central challenge by all of the profiles we interviewed. The pay gap was mentioned, especially between the managers of public healthcare facilities and university professionals. The interviewees from these groups argued that the teams are generally paid exclusively by grants from funding agencies, which makes it difficult to keep highly qualified professionals on a salary that is often lower than that offered by the market.

Interviewees from private healthcare facilities and managers from market organizations working with AI reported that the competition for professionals and the lack of resources relate to other sectors of the economy that use this type of technology and the demand from other countries. As mentioned, among the challenges to the development of AI in Brazil, job offers in AI in finance and the possibility of working remotely in foreign organizations that pay more competitive salaries appear to be significant barriers to recruiting and retaining these professionals.

For stakeholders, the turnover of professionals is a direct consequence of this challenge and has implications, such as difficulties in sustaining and continuing specific projects. It also leads to the constant need to onboard new staff, which makes it harder to achieve progress and build on existing advances.

[...] my laboratory... I'm already on my third team. I put the team together, and a bank comes along and takes everyone away. I set up the team, and a guy comes along and takes everyone. So, the workforce [...].

(ACADEMIA STAKEHOLDER)

Retaining talent is very difficult. You start, and after a year, you're at [working for a global technology company]. It's very difficult to maintain, it's very difficult to compete with the dollar, you know? There are a few who you think are good, but you can't hold on to [them].

(HEALTHCARE FACILITY STAKEHOLDER)

The training and capabilities of professionals involved in the activities necessary for AI development are also a bottleneck for the interviewees. While this challenge is noted in the Brazilian context, it is not exclusively linked to the teams developing the algorithms. According to the interviewees, especially those associated with universities, there is a significant gap in the skills of professionals responsible for building databases and using the tools. These healthcare professionals need training to record standardized information more accurately or even to interpret and analyze reports generated by machine learning. The interviewees believe that the lack of training for these professionals primarily affects the structuring and handling of data used in potential initiatives. It also has a negative impact on the ability to improve algorithm development based on the human interpretation of the results.

It's difficult to hire AI professionals. Neither the federal nor private colleges are training these professionals.

(MARKET STAKEHOLDER)

When we go into a hospital, the professional staff who interact with the data are primarily healthcare professionals or those involved in hospital management and organization. Even if there's an IT team, they tend to focus more on infrastructure than on data management aspects.[...] So, although it's nothing, if there's a team, an IT division, they're much more focused on infrastructure. [...] The human resources available focus on [patient] care. They have to be able to provide care, but there's a gap there, a very big limitation at this stage, which is you evolving and being able to introduce these things even more quickly into the routine. What people are saying about transnational medicine [...] we end up spending a lot of time getting things working better [...], because the moment I start [...], I have to organize the data. It's going to use data that also needs to be standardized. So, I think I'd say that this is the main limitation. It's even a limitation in terms of human resources because it requires.... it's an investment in infrastructure and personnel who aren't directly in the care area. It's difficult today because we're always short of resources.

(ACADEMIA STAKEHOLDER)

Interviewees from the public sector and the market also cited resistance to investing in AI as a difficulty when implementing initiatives, especially at the government level. A lack

of knowledge and mistrust of what can be developed using AI was mentioned, as was the difficulty in prioritizing the agenda since it's difficult to develop the tools in the short term and/or deal with urgent problems. This last scenario, combined with the staff shortage, led one of the public sector interviewees to discontinue their AI team and reassign it to attend to other management priorities.

This difficulty lies in developing the tools themselves (or algorithms) and in terms of interoperability. In this case, the government and the market interviewees identify both resistance and mistrust among the players about the feasibility of integrating the data without infringing any regulations. There may also be conflicts of interest, ranging from engagement and the prioritization of the issue on the agenda at different levels of the executive branch to conflicting perspectives on the ownership of the data to be used to develop the algorithms. In this regard, two accounts are worth highlighting: An interviewee from the public sector, who suggests that the most significant difficulty in achieving interoperability is the articulation between the different players responsible for the facilities that hold the data, and an interviewee from the market, who says that it is necessary to agree with clients on the transfer of data in order to develop the solution. Many organizations consider this material a strategic resource because it can guarantee, for example, ownership over algorithm development.

The network [needed to guarantee interoperability] isn't just a technological issue; it's a human issue because the major problem we're seeing isn't putting the paraphernalia in place and letting a neural network run. That's not the problem. The problem is still connecting people.

(PUBLIC SECTOR STAKEHOLDER)

The client's understanding of handing over data. I'll say it again; my client has to know that by giving their anonymized data, whatever they want to structure it, they're doing it for their own good, right? [...] We suffer a lot from this when we go to a new client. They think it's like it used to be, that I'm going to give you the intellectual capital of my professionals, and that's why I have to have a royalty share in that product. But that doesn't work with AI, because when you give me access, I'm not only going to generate this product but I'm also going to generate a series of products within the hospital. [...] So, I think this comes from the players understanding that access to data is important and that it benefits the players themselves.

(MARKET STAKEHOLDER)

Once again, regulation, specifically the LGPD, is one of the hurdles the interviewees face in implementing their initiatives.

It was referred to as a mechanism that, in principle, should guarantee the safe handling of data but which operates as a barrier to innovation. The notion of de-identifying and anonymizing how to build access to data in a structured way and guaranteeing these prerogatives are topics under discussion and open to interpretation, which may imply limitations when it comes to accessing data and developing solutions.

> [...] the LGPD concerning sensitive data is something that healthcare needs to be very well-structured around, otherwise it won't progress. If we make it too restrictive, it's going to hinder the ability to test anything.
>
> *(ACADEMIA STAKEHOLDER)*

> [...] the LGPD, which should be a mechanism for opening doors, is currently being used to lock out those who don't cooperate with us. This difficulty in accessing data is crucial for developing any kind of model, and development is not possible without it.
>
> *(ACADEMIA STAKEHOLDER)*

Finally, one of the difficulties most frequently mentioned by the interviewees is the actual capacity of the algorithms. This problem is associated with the difficulty in building data interoperability, whether due to the lack of a national database, and the lack of coordination between the players that make up the sector's ecosystem, or even the regulatory difficulties of combining information from users of healthcare systems in different facilities (private and public) in a "de-identified" way.

Many stakeholders talk about seeking partnerships with other institutions to address this challenge. Most interviewees, however, cited initiatives that rely on data that is limited to the specific facility involved — such as data on a particular pathology related to the facility's specialty or patient data from a specific hospital. In the case of governments, this may involve data from particular municipalities or states. This means that the explanatory and predictive power is constrained by the profiles that are accessible within these spaces; in other words, they are limited by the restrictions imposed by the databases available. Interviewees across all segments note this limitation, and it is one of the main risks identified for implementing AI tools in the healthcare sector, as discussed in the next section.

RISKS ASSOCIATED WITH IMPLEMENTING AI IN HEALTHCARE

Stakeholders from various fields of work point to three main risks when implementing AI tools: algorithmic bias, experimentation, and data security. How these risks are dealt with varies between the interviewees, apparently depending on the sector in which they operate, and this is especially true for data security.

As for algorithmic biases, the interviewees pointed out that this risk is intrinsically linked to access to population data. As there is no nationwide interoperability of patient data in public or private networks, algorithm learning is limited to the databases accessible to the teams responsible for the different initiatives.

The main issue for interviewees is why the algorithm only responds "about the population whose data it has accessed," which may not be an exact snapshot of the general population and may tell a fragmented story that does not have all the patient's information. This means that there is, on the one hand, a risk of bias, in which the algorithm is only able to make predictions for specific profiles but it also makes it risky to extrapolate the conclusions developed for part of the population to include the rest, thereby potentially resulting in "algorithmic racism," as an interviewee pointed out. To deal with this type of risk, the interviewees cited caution in generalizing the results and comparing the diagnosis with the specialist knowledge of professionals. However, for none of the groups, these actions seem sufficient to mitigate the risk of algorithmic bias.

In part, the stakeholders perceive that working with AI solutions involves risk because it is a time of experimentation. The interviewees point out that AI solutions are often seen in experimental phases, implying a failure risk when applied in real clinical settings. Although market players mentioned this aspect less, the interviewees observed this situation in initiatives in different fields. To mitigate this risk, they recommend monitoring the development of the algorithm and continuous validation of the results by professionals.

Finally, interviewees highlight the risks associated with data security and compliance with LGPD guidelines. All

interviewee profiles emphasized the importance of data de-identification for compliance while also mentioning the technical challenges of developing solutions based on de-identified data. This concern is particularly strong among market professionals and healthcare facility managers. These groups often view data security and privacy risks through the lens of protecting the organization. It is about ethically safeguarding user data and avoiding risks to the organization's operations. They mention actions undertaken by the company's legal departments to create guidelines and as part of the teams monitoring these initiatives.

AI AT THE FRONTLINE OF HEALTHCARE

In the second phase of the research, five interviews were conducted with healthcare professionals who use AI directly at the frontline of healthcare. This phase used a different interview script to explore issues different from those in the initial stakeholder interviews. The focus of these interviews was on specific topics such as: The practical uses of AI tools in clinical practice, the adoption process of these tools, their potential and the challenges associated with their daily use, perceptions of the new work environment, potential changes in the professional-patient relationship, and other everyday impacts experienced by healthcare professionals. The experiences shared by the interviewees in this phase provide concrete insights into the previously discussed topics, enhancing the understanding of AI's role in clinical practice.

Before delving into these topics, a brief description of the AI tools mentioned by the interviewees will be provided. Two interviewees work as nurses for a healthcare plan and interact with plan members via online chat. In this process, they use an AI tool that accesses message records, extracts information from these interactions, and summarizes the dialogue based on a predefined command that organizes critical information, such as the onset date of the symptoms and warning signs. All the information the patient provides during the chat is converted into a summary that becomes part of their medical record.

Two other people interviewed at this stage are from radiology and work in hospitals. Finally, the fifth interviewee was an intensive care nurse who is currently a university professor

who both teaches and conducts research on the use of AI in healthcare. These three professionals described various tools that are part of their daily routines.

The first is an AI system designed to automatically detect multiple sclerosis lesions (MS is a degenerative disease of the central nervous system) in patients undergoing consecutive examinations. Lesions are detected by MRI, with a follow-up conducted via repeated scans over several months. Assessing these scans, however, can be challenging for doctors and radiologists, as thousands of lesions may appear in the brain over time. Comparing successive images to identify new lesions adds significant complexity, so AI assists in this process to reduce the possibility of error.

In addition to this system, other tools were reported for disease detection, such as identifying intracranial hemorrhage in cranial CT scans, which are particularly useful in emergency care. When AI detects hemorrhage, it alerts the doctor to prioritize that report in the queue. Since hemorrhage is not a visible symptom but requires urgent attention, optimizing time may be crucial in saving the patient's life.

Sometimes, we have as many as 60 scans to review. Each MRI can take 15 to 20 minutes to analyze. With 60 exams in the queue, it's likely that a scan could be reviewed in one to two days, by which time a patient who had had a stroke might have already been discharged. This no longer happens because of the tool; it significantly speeds up stroke treatment.
(FRONTLINE HEALTHCARE PROFESSIONAL)

Some AI tools have improved the efficiency and precision of the work of professionals in radiology. The first example is a voice recognition system for transcribing reports implemented in 2013. This technology optimizes the documentation process by allowing radiologists to transcribe reports quickly and accurately, thus improving workflow. Image acquisition and reconstruction tools are also integrated into the equipment produced by major manufacturers: Software is incorporated into the devices, which optimizes image capture.

For those working with computed tomography (CT) today, we employ AI methods in the facility that allow us to use lower doses of radiation and receive scans more quickly.
(FRONTLINE HEALTHCARE PROFESSIONAL)

An AI solution for optimizing input management was also mentioned. This is software for an injection pump to reduce iodine waste. When the injection pump was being reloaded, the tool was able to identify any significant losses of the injectable material due to inadequate techniques being used. Storing and analyzing this data led to a training tool being developed that reduced losses from 10-20 ml. per fill to a maximum of 3 ml.

In addition to these tools, which are found in clinical practice or in the hospital management of those we interviewed, others were mentioned as being in the testing or development phase, such as: ML systems for predicting maternal mortality in the puerperium; the identification of chronic non-communicable diseases using a neural network methodology; and a metaverse, the simulation of environments for teaching and training purposes, in which students play characters that are in fictitious care. The aim is to make all these systems freely available in the SUS.

We also highlight the case of a university where an interviewee works. It has introduced AI tools into the basic training of healthcare professionals and formed a digital health committee to encourage the use of this technology. One of the tools designed for students is a decision-support app for use with patients on mechanical ventilation. The app gives a preliminary patient analysis and offers guidance on improving oxygenation, posture and position adjustments, aspiration, and ventilator parameters.

As in the previous examples, with the incorporation of AI in healthcare, professionals in the sector have experienced a transformation in their daily practices. AI tools play various roles, from a more efficient interpretation of images to improved patient care management. This technological advancement, however, also raises questions about the challenges and operational limits of the tools, the need for human supervision, and the ethical implications involved. In this context, healthcare professionals' perceptions of the use of AI have provided relevant elements for understanding how these innovations have been received in the clinical environment.

POSITIVE IMPACTS PERCEIVED FROM USING AI AT THE FRONTLINE OF HEALTHCARE

Overall, the interviewees reported positive experiences using AI systems to support them in their daily tasks:

It optimizes our [use of] time a lot. We used to spend a lot of time in our daily routine recording and re-reading every conversation we had had; sometimes, we had very lengthy ones in the chat. Now [AI] can summarize this a lot and optimize our use of time.

(FRONTLINE HEALTHCARE PROFESSIONAL)

Many of the doctors today can no longer work without [AI] tools. So, if a problem arises with multiple sclerosis, for example, that's not processed in the examination and now needs a medical report, people become insecure about preparing one without the tool. It doesn't mean that the tool is replacing the doctor's assessment: he's going to have to look at the two images and compare them anyway. We don't accept what the machine says as being true; the doctor's going to have to assess it. However, as the machine is extremely sensitive to detecting new lesions, if it doesn't detect one, the chance of the patient having one is almost zero. So, when the machine doesn't work, and the examination isn't processed [by AI], people miss it. It's already created a "dependency."

(FRONTLINE HEALTHCARE PROFESSIONAL)

The positive impacts of AI systems appeared in some detail in some interviews. There were mentions of the positive contributions that technology has made to: reducing the time taken screening patients or in decision-making; increasing the safety and standardization of records, medical records, and documentation; improving relationships with patients; and relieving the emotional burden that is sometimes imposed on healthcare professionals as a result of complex or numerous procedures.

I understand that AI minimizes the chance of registration errors or omissions. It provides greater security by ensuring that everything we discuss with the patient is accurately documented, including any instructions I give them.

FRONTLINE HEALTHCARE PROFESSIONAL)

It allows me to have an easier dialogue with the patient and helps me build a stronger rapport and conduct a more thorough and in-depth assessment. Instead of worrying about how much I need to summarize all that, AI provides me with a summary regardless of how much I've discussed with the patient.

(FRONTLINE HEALTHCARE PROFESSIONAL)

It's made the process much faster because we used to spend a lot of time reading and transcribing patient messages in our own words rather than actually interacting with the patient. With AI, I can now focus more on the patient rather than on registering [what they're saying], knowing that AI will provide me with a backup of all the information I've exchanged with them via chat messages.

(FRONTLINE HEALTHCARE PROFESSIONAL)

PERCEIVED LIMITATIONS AND NEW DEMANDS BASED ON THE USE OF AI TOOLS

It is important to note that although healthcare professionals see AI tools as a resource that helps their work, they have limitations and require constant supervision. As an example of a restriction, the AI tool that systematizes patient information offered on a care chatline does not process audio or read images, so Frontline healthcare professionals must interpret this data, transform it into medical language, and record it manually.

There is also a consensus among these interviewees (Frontline healthcare professionals) that AI should always be used as a shortcut and not be the final producer of a diagnosis or content. This point was highlighted many times because, as the interviewees said, generative AI tools can "hallucinate," i.e., provide answers that have no basis in reality, so a review by a responsible professional is always necessary. However, although "hallucination" is a point that requires attention, it is not enough to generate distrust among the professionals we interviewed. We identified a reasonable level of trust in the tools among the interviewees who use these technologies in their clinical practice.

I trust it because, while it can sometimes "hallucinate," it also provides me with the information I might forget. So, it has its pros and cons. I don't think we can rely on it 100%; that's why I'm still here. If AI could do everything, nurses wouldn't be needed anymore. It's good that I'm still here. But we definitely need to go through this process, as AI doesn't interpret some things in the way a healthcare professional would — it's not a healthcare professional.

(FRONTLINE HEALTHCARE PROFESSIONAL)

We always need to review the end of the report, and some terms are very medical, which must be used to train it [AI]. It sometimes writes incorrect words due to voice recognition errors, and we have to correct them and teach it. It has this learning model for voice recognition, but some words take a while for it to recognize and learn in order to correctly write a term that's very medical.

(FRONTLINE HEALTHCARE PROFESSIONAL)

In addition to initial perceptions and impacts on the work routine, other topics were questioned in the interviews, such as the potentialities, improvements, challenges, and possible risks of using AI tools in clinical practice, which are discussed below.

POTENTIALITIES, CHALLENGES, AND POSSIBLE RISKS OF USING AI IN CLINICAL PRACTICE

About the potential of this technology, professionals at the frontline of healthcare highlighted its ability to improve primary care by preventing complications arising from a patient's symptoms and reducing the number of people in the hospital who have mild symptoms. AI is believed to be useful in patient monitoring strategies in non-serious cases.

Regarding possible improvements, the interviewees pointed out that AI tools need constant adjustment and calibration. For this to happen efficiently, the healthcare professionals who use it daily must have a dialogue with the IT department or the team of developers. The tool can be constantly improved based on practical use and real demands. According to the reports, the more maintenance and calibration there is, the lower the chances of AI tools "hallucinating."

In the beginning, I had this "hallucination" problem, which happened a lot, and we always had to correct it. We'd always give them some feedback, such as changing the prompt in order to improve it to meet our demands, for example.

(FRONTLINE HEALTHCARE PROFESSIONAL)

Since we can't trust it 100% because of the "hallucination" issue, I think there may be a risk that it'll have to go through a human stage at the end. And humans are going to make a mistake because humans make mistakes. So, I think the real risk is not reviewing the AI output when it has "hallucinated." I think that's a problem.

(FRONTLINE HEALTHCARE PROFESSIONAL)

The interviews also pointed to practical challenges in implementing AI tools in the medical field, such as resistance to adopting new technologies. First, they pointed out that these applications must be integrated well into the operating systems of hospitals and healthcare facilities so that their adoption does not represent new processes and work steps. The more they are integrated into current systems and work processes, the less the chance of resistance to the technology and the better its adoption by healthcare professionals.

Another critical attention point is how easy these applications are to use and their provision of intuitive and uncomplicated interfaces. This element also influences adoption and reduces resistance to use.

I'll tell you about an experience I had when it didn't work. For example, I tried to make a website for people to access and see the outcome of the tool, but nobody used it. As I said, people need to go to the website, log in, and then see the information there, but they won't do it because it's very difficult. So, with most tools today, the process of delivering results is automatic.

(FRONTLINE HEALTHCARE PROFESSIONAL)

[It may be the] best AI tool in the world, but if you don't automate it and integrate it with your hospital system and deliver the information quickly and easily for doctors to use, it's not going to be used. Doctors won't use it if the routine is too difficult, if you add a few more steps, that's a few clicks on the screen, let's put it that way. The person's not going to use it.

(FRONTLINE HEALTHCARE PROFESSIONAL)

AI — technology — is wonderful. It's great if it's intuitive. If it's not intuitive, there's a lot of resistance to starting using the technology [...] if it increases working time, you can forget it, [because] the team's not going to do that, you know what I mean? If the time spent increases, it's an issue because doctors are paid based on productivity. So, if AI or the technology being implemented reduces their productivity, they're not going to use it. [...] If people don't receive training and go through an adaptation phase, they won't use the technology.

(FRONTLINE HEALTHCARE PROFESSIONAL)

Doctors are still hugely resistant — at least here — to using these tools. Lots of doctors are still hand-writing their notes and prescriptions.

(FRONTLINE HEALTHCARE PROFESSIONAL)

IMPLEMENTATION AND ADOPTION AT POINTS OF CARE

In addition to resistance to new technological tools, the interviews dealt with other aspects involved in their implementation and adoption by professionals at points of care. There is always a phase of adaptation at first, which can be more or less easy depending on the technical support and training offered. The informants in the survey said that training is not always available, although they believe that if the system is simple, initial instruction is sufficient, and training is unnecessary.

More training! Look, it's going to look like this. There's a key to press; you review [the case] and press the key, and it's done. There wasn't much to do; there was no training.

(FRONTLINE HEALTHCARE PROFESSIONAL)

An example of the multiple sclerosis tool is that the information reaches the doctor ready; they don't even need to know how the tool works. The information reaches them already prepared. When they go to write their report, the information is there ready for them. So, it's very easy to use it.

(FRONTLINE HEALTHCARE PROFESSIONAL)

However, depending on the tool and the circumstances, the adaptation process may not be so simple. One interviewee reported a case of a transition between two versions of an AI tool applied to radiology that did not go as planned. Instead of increasing productivity, it led to delays in the report production process.

When the version was changed, our voice recognition got worse, and the masks became more difficult, so they stayed here for another week helping us and training us. And I'll tell you that changing habits is very complicated because radiologists are very methodical and obsessive; they can't miss a single lesion. Then, you change the way the person works. It takes a while for them to adapt, even if the tool is better. They're initially resistant to working with this new technology because it's not what they're used to and it's faster. So, it was very difficult for us. In those first two months, when it was supposed to increase productivity, it actually reduced it until everyone adapted to the changes in the new version. Now that it's been six months since we changed the version, voice recognition is getting back to the way it was before.

(FRONTLINE HEALTHCARE PROFESSIONAL)

The importance of making a more significant effort when preparing healthcare professionals to work in this new scenario with AI was also emphasized. It is essential to train them not only in operating the systems but also to understand in more depth how they are built and the implications of using them. There is still a long way to go regarding transparency and explaining these solutions so that professional users understand their responsibilities and the consequences of using them.

Another important point is knowing a little about how these tools work and assessing whether this tool has been well-produced. [...] Doctors today don't need to understand programming or AI in detail. They just need to know how to evaluate the studies behind these tools — whether the dataset is well-constructed, the statistics are accurate, and the study design is sound. This allows them to assess whether the tool they might soon be using in their hospital comes from a private company or another institution and whether it's actually good and well-made. I think this is also important because many people still don't know how to evaluate an AI project or what criteria to use to determine if the work is well done.

(FRONTLINE HEALTHCARE PROFESSIONAL)

RESPONSIBILITY FOR DECISIONS AND THE DOCTOR-PATIENT RELATIONSHIP

The interviewees believe that in the event of misuse or wrong decisions based on AI-generated content, the responsibility is always human, and the determining factors for using these tools well are commitment, professional ethics, and clear guidelines for using AI.

You have to train the professionals who are going to use these tools so they're aware that they're not right all the time. They make mistakes just like human beings do, sometimes more and sometimes less. However, they make mistakes, and that means that doctors still have to assess any patient examinations.

(FRONTLINE HEALTHCARE PROFESSIONAL)

It's always the doctor's responsibility. We always say: You see the tool, but you look at the examination as if the tool didn't exist. So, the tool serves to support us, but it doesn't replace a doctor's assessment.

(FRONTLINE HEALTHCARE PROFESSIONAL)

Regarding the doctor-patient relationship, interviewees' perception is that AI tools are beneficial because, by optimizing the healthcare professional's time, there is a greater chance of quality interaction with the patient.

We can devote more time to the patient than concentrating on filling out digital paperwork. We were very concerned about this: "I have to register it in a complete way. I must pass on the information. It has to use that methodology. I'm going to spend a lot of time [doing it]. I'm seeing three people at the same time." I'm talking to three individuals, and I have to guarantee the quality of my services, of course. So, when I take a bit of that mental load and time away from something bureaucratic — from a bureaucratic step in my work — I can concentrate more on patient care, that's for sure. I can guarantee more time reading his medical records, getting to know more about him, and understanding his clinical history better so I can see him at that moment. So, I think that, yes, maybe the patient doesn't know, but behind the scenes, we're managing to improve things a lot for him.

(FRONTLINE HEALTHCARE PROFESSIONAL)

The interviews reveal that patients are not always aware that an AI tool supports healthcare professionals in their care or in analyzing their data. According to the interviewees, in most cases, healthcare facilities do not report, notify, or consult patients on this issue, even though they are AI healthcare tools' ultimate beneficiaries. As most of the tools in use today are geared more towards internal processes, such as triage,

registration, or medical records, the professionals interviewed believe this technology does not directly interfere with care, nor do they consider it necessary to inform patients about its use.

How would they know if I'm using AI or not in all these methods? I don't think they do. They don't know whether I'm using voice recognition or a technique to reduce radiation. They might know that I have more advanced equipment, but they don't have that precise knowledge.
(FRONTLINE HEALTHCARE PROFESSIONAL)

The patients themselves aren't aware that these tools are working. What matters is the impact on treatment and the speed of treatment, such as the early detection of a stroke or hemorrhage. Patients with these conditions are treated more quickly here at our hospital.
(FRONTLINE HEALTHCARE PROFESSIONAL)

Although the lack of notification and consultation practices regarding the application of AI tools was not a significant issue of concern for the professionals we interviewed, one of them gave a more in-depth reflection, which points to a possible change in the future scenario. This can happen through more active patient participation in healthcare as an expression of greater ownership of the use of technology in their lives. To this end, the perspective of the end beneficiary should be considered in clinical care and also in the development and implementation of technological tools.

He can't be a person who just absorbs [AI]. There'll be those who absorb more. [...] He needs to be someone [looking after] his own health; he starts by demanding better quality. He's not a passive being who just stands there waiting for the result. He knows his body, he knows health, he knows where a lot of things are getting in the way, he needs to be listened to, and the technologies need to have this framework.
(FRONTLINE HEALTHCARE PROFESSIONAL)

The individuals interviewed say that the patient data used by these tools is protected, and hospital ethics committees approve their practices. According to the reports, before AI processes, all images and information go through an anonymization process to guarantee identity protection. No personal information is used, and the tests are not used for the machine to rework, reinforcing the care taken when handling the data. Furthermore, in all the experiences reported, the databases are private and not shared, ensuring the integrity and confidentiality of the data of the patients involved.

FUTURE PROSPECTS FOR HEALTHCARE PROFESSIONALS

A final topic addressed in the interviews with frontline healthcare professionals dealt with their future prospects, considering the use of AI technology in clinical practice and at points of care.

In one of the interviews, a medical professional thinking about the future of professionals highlighted the importance of a curricular reform that prioritizes the theme of technology in healthcare. He believes this change is essential to reduce the resistance of healthcare professionals to incorporate AI into their daily lives. He argued that the sooner professionals are prepared, the quicker they will understand the benefits of AI in their work. This approach would also help normalize their views on the subject and dispel any fears of human beings being entirely replaced by machines with the advent of AI technologies in healthcare.

Nursing professionals who include AI in their daily practices do not believe they will be replaced, and they highlight two essential factors: The need for human empathy in patient interactions and clinical experience.

I'm not entirely replaceable. There are things I do that I'd really like to see automated because they involve very bureaucratic tasks that consume a lot of my time. No one who studies nursing or medicine does so to fill out forms; that's a fact. When we have to do these tasks, we become tired of that stage because we studied to care for people, not to handle bureaucratic steps that a machine could manage.

(FRONTLINE HEALTHCARE PROFESSIONAL)

I'm sure technology won't solve everything, but at the same time, professionals need to be more open, dynamic, and prepared to embrace a new era. It's already here; we`ve been overtaken by it. So, we need to wake up and see where we can contribute, how it benefits us, and what the impact on our profession is. Otherwise, we risk being "run over by a tractor."

(FRONTLINE HEALTHCARE PROFESSIONAL)

This stage in the research revealed that AI brings significant benefits to the working practices of frontline healthcare professionals, such as optimizing their time, reducing mental workload, and improving the professional-patient relationship. However, operational challenges, costs, and the need for human

oversight highlight the complexity of implementing AI in healthcare. Therefore, issues of ethics, medical accountability, and consideration of the patient's perspectives are areas that are still underexplored in the adoption of AI in healthcare.

FINAL NOTES

The qualitative survey collected the perceptions and experiences of strategic players in the healthcare sector working with AI in different spheres: Academia, the public sector, the market, and healthcare facilities. We aimed to understand the state of AI initiatives in healthcare in Brazil and identify themes, concerns, and expectations on the public agenda, intending to provide an overview of this topic in the country.

The interviews revealed that the country is at an early stage in developing and implementing AI tools in healthcare. There is a climate of optimism and great expectation regarding the potential of AI, especially in the private sector, where development is more advanced than in the public sector. The main advances mentioned include using AI to improve healthcare and management processes, emphasizing increasing efficiency and reducing costs.

Despite their enthusiasm, the study's interviewees recognize that the country is just starting on this journey and faces challenges, such as the lack of a national strategy, regulatory issues, and data quality and integration deficiencies. The evolution of AI in Brazilian healthcare is perceived as uneven, with the private sector leading the movement while the public sector faces bureaucratic challenges and a lack of resources. The main gaps pointed out relate to the quality of the data and the lack of specific regulations for the use of AI in healthcare. According to the survey respondents, it is essential to tackle these challenges to make significant progress in the field of AI applied to healthcare in Brazil.

Concerning the potential of AI, three main areas of opportunity were highlighted: improvements for patients, healthcare professionals, and service providers. AI can potentially increase access to healthcare services and improve diagnostic accuracy for patients. In contrast, for healthcare professionals, AI tools can reduce bureaucratic processes, optimize time spent on administrative tasks, and support clinical

decision-making, thus promoting faster and more accurate diagnoses. They can also increase the capacity for care in regions with a shortage of specialists. For healthcare service providers, AI can lead to operational efficiency, optimize resources, and improve management, logistics, and service processes in both the public and private sectors.

All this potential can also be used to tackle the inequalities found in healthcare provision by expanding access to services and reducing the impact of scarce resources in the country's most disadvantaged regions. In summary, AI is perceived as a tool that can promote significant improvements in various aspects of healthcare in Brazil, from operational management to clinical care, with benefits for healthcare professionals, patients, and society as a whole.

Still, regarding opportunities, the SUS was pointed out as the most significant Brazilian differential in obtaining a sufficiently large volume of diverse data to develop robust and reliable algorithms and applications. In the view of those interviewed, the RNDS has enormous potential to promote interoperability and create an environment that is conducive to the development and application of AI in the Brazilian healthcare sector. To do so, the data needs to be of high quality and not just in large volume, which implies the importance of protocols and routines that, among other things, ensure the standardization of health information.

In this climate of optimism about AI in healthcare, the interviewees found it difficult to respond to the possible risks of using AI. However, concerns were raised about the privacy and security of patient data. The risk of sensitive information being leaked was pointed out as a relevant issue, considering the possibility of the data used and/or generated by AI tools being employed inappropriately. The discussion on the rights of the users of healthcare systems mentioned the limitations of AI, emphasizing the need for digital education to promote understanding of the benefits of technology. Risks related to ethics were also mentioned, although this issue was not as prominent in the discourse. According to the interviewees, in order to mitigate the potential risks associated with the advance of AI in healthcare, there needs to be transparency in the decisions that underpin the construction of the tools, a clear

governance structure for risk assessment, and the development of bias-free algorithms that produce reliable results.

Still, on potential risks, the responsibility for possible AI mistakes lies, by consensus, with the professionals who use the tool in their work practice. The research revealed, therefore, the difficulty of elaborating on social and collective mechanisms of an ethical order, with the prevalent perception being that of individual accountability when faced with AI-based decisions.

The study also mapped ongoing AI actions, initiatives, and projects in healthcare in the country in the four stakeholder segments that were interviewed. Reports of initiatives that have taken place in the academic sphere and the public and private sectors were collected to present a broad overview of the current scenario in Brazil. In general, the AI initiatives applied to healthcare that are underway in Brazil aim to optimize management processes and improve the quality of care. Each sphere has its specific focus: in academia, the focus is on developing algorithms emphasizing Brazil's population diversity and correcting algorithmic bias, as it can have harmful social implications. The market focuses on generative AI to create various products, such as chatbots for patient care and tools for optimizing imaging diagnostics. In the public sector, the focus is on digitalization and the interoperability of data. At the same time, in healthcare facilities, there is an emphasis on various projects to help with diagnosis and management. The similarities and divergences observed between the initiatives of the different segments we investigated illustrate the potential of the tools and the difficulties encountered in developing and implementing them.

Finally, to complement the topics discussed, the study focused on the experiences of five frontline healthcare professionals who use AI solutions in their daily tasks. The interviews explored the practical uses of AI tools in clinical practice, the adoption process, the potential benefits, challenges, and the daily impacts these professionals experience. They reported positive experiences and specifically singled out the optimization of their time, a reduced mental workload, and consequently improved relationships with patients. However, they also noted challenges in the effective implementation of

AI tools, including resistance to their adoption by healthcare professionals due to difficulties with adaptation, inadequate training, and the need to integrate AI solutions with existing healthcare systems.

There was an emphasis on the importance of human supervision in the adoption of AI solutions, the need for continuous calibration of the tools used, and the ethical responsibility of healthcare professionals regarding technology-supported diagnoses. Looking to the future, the Frontline healthcare respondents highlighted the need for curricular reforms to include health technology in professional training, thus preparing them to use AI. They also reflected on the significance of empathy in patient interactions and valuing the clinical experience of professionals because they believe human care will always be imperative concerning technology.

REFERENCES

Adler-Milstein, J., Aggarwal, N., Ahmed, M., Castner, J., Evans, B. J., Gonzalez, A. A., James, A. C., Lin, S., Mandl, K. D., Matheny, M. E., Sendak, M. P., Shachar, C., & Williams, A. (2022). Meeting the moment: Addressing barriers and facilitating clinical adoption of Artificial Intelligence in medical diagnosis. *NAM Perspectives, 2022*. https://www.gao.gov/products/gao-22-104629

Brazilian Academy of Sciences. (2023). *Recomendações para o avanço da Inteligência Artificial no Brasil: GT-IA da Academia Brasileira de Ciências*. https://www.abc.org.br/wp-content/uploads/2023/11/recomendacoes-para-o-avanco-da-inteligencia-artificial-no-brasil-abc-novembro-2023-GT-IA. pdf

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial Intelligence, bias and clinical safety. *BMJ Quality and Safety*, *28*(231-237). https://qualitysafety.bmj.com/content/28/3/231

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. It addresses the processing of personal data, including in digital formats, by natural or legal persons of public or private law, with the aim of protecting fundamental rights to freedom and privacy and the free development of the personality of the individual. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Ministry of Health. (n.d.). Rede Nacional de Dados em Saúde. https://www.gov.br/saude/pt-br/composicao/seidigi/rnds

Ministry of Health. (2020). *Estratégia de Saúde Digital para o Brasil*. https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf

CONCLUSIONS
Public policy drivers for using Artificial Intelligence in healthcare

Glauco Arbix[1] *and João Paulo Veiga*[2]

1 Professor at the Department of Sociology of the University of São Paulo (USP).
2 Professor of Political Science and International Relations at USP and a researcher in the Center for Artificial Intelligence (C4AI).


010
101
010
010
101
010
010
010
010
010
010

INTRODUCTION

The healthcare sector in Brazil is an economically vital area. In addition to accounting for 9% of Brazil's Gross Domestic Product (GDP) and generating around 9 million direct jobs, healthcare offers a considerable opportunity for the country to align with the new technological cycle, which has at its heart one of the most powerful technologies that humanity has ever created: Artificial Intelligence (AI).

AI is profoundly transformative in research, the economy, and society in general. Because of its flexibility and power, it is becoming increasingly established as a general-purpose technology like automobiles, electricity, television, computers, and the Internet. Its diverse and multifaceted characteristics make integrating it with other disciplines and tools easy, which supports its penetration into practically every area of the economy and society.

Over and above its reach, however, AI has become essential for generating innovation. In healthcare, it has either led or shared in enormous advances in biochemistry, biology, genetics, materials, energy, and all those domains that shape and inform modern care for the population. The same is true in management and the integration between science and medical infrastructure, in addition to AI potentially increasing transparency for the users of healthcare services. The same happens in advanced research into new drugs and medicines, in manufacturing processes, in equipment, and in improving the quality and efficiency of the services needed to guarantee a basic right of the population: the right to health care. Because of all this potential, it has become practically impossible to address healthcare strategies without emphasizing and focusing on the impacts of the new AI technologies.

The central concern that inspired suggestions for developing a public health strategy, as presented in this chapter, was informed by the need for an agenda of changes in the Unified Health System (SUS), starting with a significant expansion of sustainable access to those public and private services that look after the well-being of the population.

If AI technologies are used ethically and responsibly, they can lead to the implementation of more efficient healthcare

policies and the raising of the level of access to healthcare services, especially for the most vulnerable. This horizon becomes especially relevant in a country like Brazil, which is marked by historical deficiencies and profound social inequalities, which are expressed in income, regional disparities, and the current limitations of the SUS when it comes to offering better quality care for a broad spectrum of medical procedures that range from basic to complex.[3]

AI EXPERIENCES IN HEALTHCARE

In line with global trends, more than 30 million medical consultations were carried out remotely in Brazil in 2023 alone, according to data from the National Federation of Supplementary Health (FenaSaúde, 2023), which is in stark contrast to the 11 million remote consultations that were carried out between 2020 and the end of 2022.[4] The Ministry of Health is expecting more than 50 million consultations in 2024 (Folha de São Paulo, 2024).

In addition to medicine, telehealth includes remote care in several areas, such as nursing, physiotherapy, and psychology. According to established Brazilian rules, healthcare professionals have the autonomy to decide whether they use this practice, including the first consultation. The legislation also guarantees the right of the patient to refuse remote care and the confidentiality of their data. According to the Ministry of Health (MS), in 2023 alone, 1,200 municipalities used remote electrocardiogram services, with an average of 6,000 medical reports being issued daily. The MS's plans to expand the digitalization of healthcare have included several telehealth procedures, such as mechanisms that improve patients' access, the monitoring, and continuity of their care, and the management of waiting lists. These resources do not replace all types of medical care. Yet, they complement face-to-face care, leading to significant gains in access to

3 Some of the recommendations for this work were based on analyses found in the *São Paulo Healthcare Proposal 2022* (*Proposta Saúde São Paulo 2022*) document, organized by SindHosp (2022) and based on technical consultancy work carried out by the Brazilian Center of Analysis and Planning (CEBRAP).

4 Despite this increase and its practicality, remote care was only regulated by the Federal Council of Medicine (CFM) in 2022 (Law No. 14,510/2022), because telehealth had been allowed on an emergency basis during the COVID-19 pandemic (Law 13,989/2020).

primary care, monitoring patients with chronic diseases, care in remote areas that still have precarious services, and cost reductions and convenience. Given these advantages, the number of clinics and hospitals using telehealth has increased rapidly, including management, planning, predicting the risks of hospitalization, and collecting and delivering material for examination.

Telehealth systems have shown to be increasingly dependent on AI technologies. This synergy has expanded the scope of healthcare services by including more complex procedures, such as image analysis and the interrelationship between symptoms and the biomarkers of clinical data to characterize and forecast diseases, enable the use of sensors, promote surgery, and monitor patients (Joshi et al., 2021; Liu et al., 2022; Weenk et al., 2020). Incorporating AI resources and methodologies has boosted the use of apps for measuring vital signs, detecting movement, and even recognizing the cognitive parameters that can indicate confusion, falls, or psychological changes (Shaik et al., 2022). Many algorithms have leveraged a mass of exploratory activities in different countries and regions seeking to change the volume and quality of healthcare services, which has led to significant gains for the population, especially the underprivileged.

This same momentum is observed in medical research, both in genome sequencing and in the development of new drugs and treatments that are enabled by machine-learning (ML) techniques. Doctors and scientists have been increasingly helped by artificial agents when they need to interpret large volumes of data in a short space of time (Helm et al., 2020; Krittanawong et al., 2022) to predict an early deterioration in the health of patients living with chronic disorders (Guo et al., 2022; Liu et al., 2022). AI systems are increasingly able to process the data that recognize and identify patterns that help healthcare professionals with their decision-making (Dean et al., 2022). In fact, the increase in computing capacity and speed in recent years has led to the development of artificial neural networks and deep learning (DL) algorithms that move highly complex databases (Kalfa et al., 2020) and automate and control tasks to avoid human error and increase patient safety (Tandel et al., 2022).

A study based on the Web of Science, Scopus, Springer, Pub Med, Science Directy, and ACM Digital Library databases (Shaik et al., 2023) identified a wide range of health domains in which AI has boosted telemedicine and had an impact on both advanced and emerging countries. The research results indicated that AI-driven telemedicine has progressed rapidly in the United States (USA) and China since the pandemic, a trend that has also been seen in Brazil. A 2023 survey by the Kaiser Family Foundation showed that almost all health plans in the US offered telemedicine procedures in 2022, while telemedicine in Japan became part of the public health system and has its own strategy with a focus on AI. The Japanese Ministry of Health has implemented remote medical services in more than two thousand hospitals and clinics nationwide to treat chronic diseases, support emergency care, and conduct remote consultations in distant regions. Telemedicine in the United Kingdom is an integral part of the National Health Service (NHS) – the British health system that inspired Brazil's SUS –which chose AI to analyze patient data, prepare preliminary diagnoses, and enable mechanisms to prevent hospitalization. In this same vein, the World Health Organization (WHO) indicated in 2024 that AI technologies were active in telehealth systems in all European countries (WHO, 2024).

RISKS OF AI IN HEALTHCARE

This rapid advance of AI in healthcare is not without its technical and ethical risks. From a legal and protection of society perspective, regulating the use of data is essential. This starts with the confidentiality and protection of the data of those who use healthcare services, which are recognized as two distinct fundamental rights for individuals' protection, autonomy, and dignity. These two rights are enshrined in general data protection regulations, such as those in the European Union (EU) and Brazil. This means that the tools that process healthcare data must comply with legal requirements in order to guarantee their lawfulness, transparency, purpose limitation, accuracy, storage limitations, integrity, confidentiality, and the limitations that arise from their governance.

Healthcare is a special field in which ethical guidelines and safeguards are essential because healthcare impacts the

lives of patients and their families. This means that efforts to regulate and standardize data use and draft laws must be based on the ethical principles governing medicine. The challenge, however, is to remodel ethical references to account for the new, interactive, and digitalized processes carried out via a smartphone, tablet, or computer and configured in a radically different way than the medicine has used for centuries. It is no wonder that patient safety, transparency, and the explainability of the procedures used are essential for making users and professionals accept and incorporate the new digitalized processes into medical practice. However, this is not always easy, given the nature of the new resources.

Many algorithms are more efficient than humans in processing and interpreting complex data and predicting results. They are, however, unable to demonstrate how these conclusions were reached or if there were flaws in their statistical path. The structure of deep neural networks is not self-explanatory: it is highly opaque and expresses a dynamic known as a black box (Yang et al., 2022). Therefore, it is impossible to predict what neural networks learn from the data, but from then on, they can mobilize the resources required for discriminating patient information (Chen et al., 2021).

This uncertainty, which becomes greater when there are imbalances and failures in the formation of the databases, compromises the rapid adoption of these technologies in healthcare. In other words, even though it is widely known that AI has profoundly transformed applications in the healthcare field, and there have been significant advances in telehealth, challenges such as adequate data processing, explainability, and privacy need to be addressed to reduce uncertainty and increase reliability.

Although they are still in the early stage and limited, consistent experiments are being conducted to increase the transparency of AI systems by indicating the path they took in reaching their decisions. Such experiments seek to increase the confidence of doctors and healthcare professionals by not over-emphasizing the technical attributes of the systems (Lauritsen et al., 2020). The search for reliability criteria also forms the backdrop when doctors are prompted to move away from conventional diagnoses based on population averages

and to try to consider the individual variability of patients and their responses to treatment. In this sense, the link between telemedicine and AI resources allows for personalized monitoring that focuses on the patient's clinical life, which has proven to be a handy tool in the treatment of chronic diseases such as mental health disorders, diabetes, heart disease, and others (Mukherjee et al., 2020). Despite being promising, these procedures need to be anchored in AI platforms and stored in the cloud so the data can be analyzed. This immediately raises privacy and security concerns about this health data and the high cost because storage requires enormous technological resources and consumes large amounts of energy.

Despite these risks, medical research increasingly uses AI due to its advantages compared to traditional scientific practices and diagnostic, clinical, and surgical procedures. The main goal of researchers and the experiments is to build a personalized and reliable picture of patients that can capture signs of decline in their health conditions as early as possible and, consequently, predict adverse events and improve the accuracy of the treatment and its results.[5]

The works cited in the previous chapters of this book provide a small sample of both the vigor of ongoing research and the caution needed if AI is to become established as a powerful instrument in everyday medical practice. The reliability of the procedures, resources, and tools is one of the greatest challenges – if not the greatest – that research in medicine currently faces (Mohanty & Mishra, 2022).

The benefits AI can bring to healthcare do not erase the highly complex nature of this technology. Therefore, institutions must be committed to the responsible use of AI to minimize risks and maximize opportunities. The WHO (2021) has released guidelines on the ethical use of AI that can be summarized in six points: (a) Humans must control health systems and make any decisions relating to health; this should not be done exclusively by AI; (b) Developers and those responsible for AI systems must monitor and ensure the full

5 Particularly in the monitoring of patients with chronic illnesses there is a profusion of algorithms and wearables, which also prescribe the dose of the medication and the time it should be taken (Watts et al., 2020; Wu et al., 2022; Yu et al., 2023; Zheng et al., 2021).

functioning of all the tools, in order to guarantee compliance with all safety rules; (c) Developers are also responsible for publishing data and information about products and how they are handled in a fully transparent manner; (d) Health systems that adopt AI should ensure adequate training for the professionals responsible for the tools they use; (e) To promote diversity and avoid biased algorithms, AI training should use data taken from different nationalities, genders, and ethnicities; (f) AI tools should be constantly evaluated and, based on their performance, improved.

Despite the WHO's guidelines, new digital technologies are multiplying much faster than the debate and the efforts to regulate and discipline health agents by establishing rules to limit their responsible and ethical use. The relationship between patients and doctors is multidimensional and, when healthy, is based on transparency about the procedures used. This concern must be redoubled when AI is involved, given the novelty of the technology and the insecurity it still generates. Responsible medical practice must also distance itself as far as possible from any technological determinism that gives rise to uncritical views. It must focus on improving the quality of the interactions between doctors and patients, which defines what is properly human and what can be mapped out and codified by statistics, but that is difficult to understand and sense artificially. In other words, AI must be structured so that it works with doctors to benefit end users and does not replace healthcare professionals.

WELL-PREPARED PROFESSIONALS AND THE DIGITALIZATION OF HEALTHCARE IN BRAZIL

Large private corporations worldwide are investing heavily in AI applications and research in healthcare, especially those based in the US, China, Germany, Japan, the UK, and developing countries such as India. Notable research centers include Stanford and Harvard Universities, the Massachusetts Institute of Technology (MIT) in the US; the Universities of Oxford and Cambridge and the Allan Turing Institute in the UK; and the Universities of Beijing and Tsinghua in China. In Brazil, the University of São Paulo (USP), the State University of Campinas (Unicamp), and the Federal Universities of

Minas Gerais (UFMG), Rio Grande do Sul (UFRGS), and Rio de Janeiro (UFRJ) are among those most dedicated to this field of research.

Brazil has guidelines that were defined by the Ministry of Science, Technology and Innovation (MCTI) in 2021, and that go to make up the Brazilian AI Strategy (EBIA) (MCTI, 2021; MCTI Ordinance No. 4,979/2021). Eleven research centers are also in the process of being set up, which are funded by the federal government, state research support foundations (such as the São Paulo State Research Funding Agency [FAPESP]), private companies, and the Brazilian Internet Steering Committee (CGI.br). Bills are also being widely discussed in the National Congress and Brazilian society to provide Brazil with its own regulatory framework for AI.

The EBIA currently provides the public sector with guidance regarding its efforts to support AI and, based on recommendations from the Organization for Economic Cooperation and Development (OECD), has established three thematic (transversal) axes and six vertical axes.

The thematic axes are:

1. Legislation, regulation, and ethical use: Dealing with the legal, regulatory, and ethical parameters for developing AI.
2. AI governance: Establishing a governance structure that promotes methods and procedures for ensuring observance of the principles of AI when developing solutions that use this technology.
3. International aspects: Dealing with cooperation and integration platforms for exchanging information, experiences, regulations, and best practices in the conduct of AI worldwide.

The vertical axes that define the priority areas for the development of AI are:

1. Education;
2. Workforce and training;
3. Research, development, innovation, and entrepreneurship;
4. Applications in production sectors;
5. Applications in government;
6. Public security.

With these general guidelines, the EBIA seeks to bring together the government, the private sector, universities, and the third sector. This broad-spectrum link in healthcare is essential, so AI technologies can strengthen, integrate, and make the SUS — a repository of large databases — more effective, and boost results-oriented research. This is a core priority in a country that needs to act urgently to overcome its history of inequality and social deficit.

The keywords that summarize the most relevant dimensions that need to become more dynamic in the SUS are: Training professionals, digitalization, integration, efficiency, quality, and funding. While this chapter deals with new digital technologies, especially AI, the authors nevertheless recognize that technology is directed and determined by society, not vice versa.

Expanding access, especially for the neediest, collecting and processing the data that are not always captured at the base of the social pyramid, and strengthening and integrating the primary care network are essential for improving efficiency in the SUS. Technologies can substantially help increase the quality of services, provide faster and more effective care, curtail the indiscriminate use of medication, and reduce premature hospitalization. Telehealth can operate in all these dimensions and increase their resolution. The breadth of the dimensions that are impacted by improving management means that any efforts to digitalize the SUS and build AI-powered telehealth systems must be addressed in an integrated manner with a national strategy for strengthening and reconfiguring the SUS, which must include innovation in the equipment it uses and in pharmaceuticals.

The COVID-19 pandemic exposed several weaknesses in the health complex in medicines, vaccines, tests, equipment, and even the basic products it uses. In other words, Brazil must promote synergies between research and the industrial complex because this is key to its health system.

The urgency of this strategy is linked to the speed with which the new technological cycle that is driven by AI and is shaking the world is developing. The specialists are practically in full agreement that the lack of public policies dealing with health innovation makes it difficult for the pharmaceutical,

medicine, and new equipment industries to evolve. They can only overcome this time lag and their deficiencies when there is coordinated action between the private and public sectors: Between companies and the Ministry of Health, the Ministry of Development, Industry, Commerce and Services (MDIC), the MCTI, and development agencies. Public policies that bring industry and research bodies closer together and provide basic inputs can reduce the weight of exports in Brazil's trade balance and free up resources to be allocated to the SUS. This means there is no way of avoiding using new technologies, especially AI when implementing a national healthcare strategy, whether in services or in industry.

DIGITALIZATION AS STRATEGY

The SUS is a conglomerate of public, private, and semi-public institutions – whether governmental or not – that brings together a world of competencies that are not always complementary and involve different management and command structures that are national, regional, state, and municipal. The complexity of this system, with its diversity, regulations, funding, training, and hiring of professionals and services, assistance, evaluation, and absorption of technology, poses enormous challenges for managing it, for the quality of the services it offers, and for the agility and efficiency of the care it provides. International experiences involving the integration of services make it possible to estimate the impact that AI can have on the control and management of the SUS in order to: (a) enable the computerization of more than 50,000 family health teams, which are spread among most of the municipalities in Brazil, and that operate in more than 35,000 Basic Health Units (UBS); (b) Boost digital resources to integrate community healthcare agents (ACS) better, and thus facilitate the presence and remote training of healthcare professionals; (c) Introduce and multiply telediagnosis procedures; and (d) Help structure multidisciplinary teams of healthcare professionals who work remotely and meet the demands of the UBS.

Information reliability has become enormously important for both services and the industry. During the COVID-19 pandemic, localized and sectoral experiments gathered data on hospitalizations and tests. However, the health system's

daily routine is different because information is collected in a fragmented manner via the SUS Information Technology Department (DATASUS), the Brazilian Supplementary Health Agency (ANS), and from state and municipal agencies. To give just one example: In 2021, the Brazilian Senate approved a bill to create a digital platform capable of unifying patients' medical records from public and private networks country-wide (Agência Senado, 2021). The *SUS* was tasked with centralizing this information, including data on prescriptions, referrals, medical records, and test reports. The law also obliged the SUS to adapt its platform to meet the requirements of the General Data Protection Law (LGPD) (Law No. 13,709/2018), with the Ministry of Health having already announced a version of the Conecte SUS system that includes electronic medical records. In 2021, the update of the Policy on Health Information and Informatics (PNIIS) (Resolution No. 659/2021) and the Digital Health Strategy 2020-2028 (DHS) (MS, 2020) reaffirmed this direction by defining the National Health Data Network (RNDS) as an environment for interoperability and communication between stakeholders in the SUS.

Despite the legislation being favorable to data integration and sharing, the obstacles to its implementation remain enormous. They include the availability of an infrastructure for this network to become a reality, the willingness of SUS stakeholders to share their data, and a strengthening of the regulatory and legal bases of the system, including ethical requirements regarding the use of patient information and costs. From a technological point of view, interoperability needs AI as a tool in order to achieve this integration.

Steps in this direction are essential for improving the management of the health system, starting with preparing anonymized databases that enable aggregated analyses that can generate knowledge about the population's health. These databases should make it possible to analyze health indicators and the outcomes of the procedures performed within a given SUS region. This information is essential for prevention, raising awareness, optimizing investment in training, and a more appropriate infrastructure allocation. With this data, the SUS would be able to monitor regional trajectories, make comparisons and projections, and strengthen

all planning activities in an unprecedented way, with greater rationalization and rigor in the use of its resources and in controlling the healthcare policy objectives.

RECOMMENDATIONS FOR FORMULATING A STRATEGY FOR AI IN HEALTHCARE

The following are suggestions for establishing public healthcare policies that use AI as the protagonist, facilitator, and enabler of innovation:

1. Prioritize AI technologies in telehealth to expand sustainable access and primary care and integrate them with moderate and highly complex medical and hospital procedures.
2. Introduce advanced management systems with data integration in the SUS that collect information in real time.
3. Encourage governance models for integrating databases that favor an advance in sharing electronic medical records.
4. Educate, train, and requalify healthcare professionals, including by remote means, so they can accompany the changes and take advantage of the opportunities offered by technology to improve the quality of the SUS and the work it does.
5. Establish safeguards for protecting the population by way of ethical regulations and specific legislation for AI.
6. Encourage research into AI applications in healthcare, whether to improve and expand care for the population or improve public policy management and quality.
7. Support advanced research in AI in healthcare and the links between universities and companies for carrying out measurable projects, monitoring, and ethically transparent projects that can offer the population tangible results.
8. Boost innovation in healthcare: Using the State's purchasing power can leverage the transfer of technology and finance large projects based on the mobilization of skills in different research areas, whether in universities, companies, or through the coordination of research centers, starting with AI institutes.
9. Consider building an AI observatory model in healthcare

to capture international trends, monitor the evolution of smart technologies in Brazil, and develop technologies to evaluate the performance of public projects and initiatives.

10. Ensure the SUS and health policies follow ethical rules based on transparency and the reliability of technological systems in order to protect the population and maintain the dignity of the relationship between doctors and patients.

REFERENCES

Agência Senado. (2021). Aprovada criação de plataforma para unificar dados do SUS e da rede privada. *Senado Notícias*. https://www12.senado.leg.br/noticias/materias/2021/05/18/aprovada-criacao-de-plataforma-para-unificar-dados-do-sus-e-da-rede-privada

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition. *ACM Computing Surveys, 54*(4), 1-40 https://dl.acm.org/doi/10.1145/3447744

Dean, N. C., Vines, C. G., Carr, J. R., Rubin, J. G., Webb, B. J., Jacobs, J. R., Butler, A. M., Lee, J., Jephson, A. R., Jenson, N., Walker, M., Brown, S. M., Irvin, J. A., Lungren, M. P., & Allen, T. L. (2022). A pragmatic, stepped wedge, cluster-controlled clinical trial of real-time pneumonia clinical decision support. *American Journal of Respiratory and Critical Care Medicine, 205*(11), 1330-1336. https://doi.org/10.1164/rccm.202109-2092OC

Federação Nacional de Saúde Suplementar. (2023). *Relatório anual de atividades*. https://fenasaude.org.br/publicacoes/relatorio-anual-de-atividades-fena-saude

Folha de São Paulo. (2024). *Atendimentos por telemedicina no país crescem 172% em 2023 após lei que regulamenta saúde digital*. https://www1.folha.uol.com.br/equilibrioesaude/2024/04/atendimentos-por-telemedicina-no-pais-crescem-172-em-2023-apos-lei-que-regulamenta-saude-digital.shtml

General Data Protection Law (LGPD). (2018). Law No. 13,709, of August 14, 2018. This law addresses the processing of personal data, including in digital media, by natural persons or legal entities, whether public or private, with the aim of protecting the fundamental rights of freedom and privacy, and the free development of the personality of the natural person. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Guo, J., Huang, X., Dou, L., Yan, M., Shen, T., Tang, W., & Li, J. (2022). Aging and aging-related diseases: From molecular mechanisms to interventions and treatments. *Signal Transduction and Targeted Therapy,7*(391). https://doi.org/10.1038/s41392-022-01251-0

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine learning and Artificial Intelligence: Definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine, 13*(1), 69-76. https://doi.org/10.1007/s12178-020-09600-8

Joshi, M., Archer, S., Morbi, A., Arora, S., Kwasnicki, R., Ashrafian, H., Khan, S., Cooke, G., & Darzi, A. (2021). Short-term wearable sensors for in-hospital medical and surgical patients: Mixed methods analysis of patient perspectives. *JMIR Perioperative Medicine, 4*(1), e18836. https://doi.org/10.2196/18836

Kalfa, D., Agrawal, S., Goldshtrom, N., LaPar, D., & Bacha, E. (2020). Wireless monitoring and artificial intelligence: A bright future in cardiothoracic surgery. *The Journal of Thoracic and Cardiovascular Surgery, 160*(3), 809-812. https://doi.org/10.1016/j.jtcvs.2019.08.141

Krittanawong, C., Johnson, K. W., Choi, E., Kaplin, S., Venner, E., Murugan, M., Wang, Z., Glicksberg, B. S., Amos, C. I., Schatz, M. C., & Tang, W. W. (2022). Artificial intelligence and cardiovascular genetics. *Life, 12*(2), 279-307. https://doi.org/10.3390/life12020279

Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Explainable Artificial Intelligence model to predict acute critical illness from electronic health records. *Nature Communications, 11*(3852), 1-11. https://doi.org/10.1038/s41467-020-17431-x

Law No. 14.510, of December 27, 2022. (2022). Amends Law No. 8.080, of September 19, 1990, to authorize and regulate the practice of telehealth throughout the national territory, and Law No. 13.146, of July 6, 2015; and revokes Law No. 13,989, of April 15, 2020.. https://www.in.gov.br/web/dou/-/lei-n-14.510-de27-de-dezembro-de-2022-454029572

Law No. 13,989, of April 15, 2020. (2020). Provides for the use of telemedicine during the crisis caused by the novel coronavirus (SARS-CoV-2); revoked by Law No. 14,510 of 2022.. https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/Lei/L13989.htm

Liu, H., Wang, L., Lin, G., & Feng, Y. (2022). Recent progress in the fabrication of flexible materials for wearable sensors. *Biomaterials Science, 10*(3), 614-632. https://doi.org/10.1039/d1bm01136g

Ministry of Science, Technology, and Innovation. (2021). *Estratégia Brasileira de Inteligência Artificial (EBIA).* https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-documento_referencia_4-979_2021.pdf

Ministry of Health. (2020). *Estratégia de Saúde Digital para o Brasil 2020-2028*. https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf

Mohanty, A., & Mishra, S. (2022). A comprehensive study of explainable Artificial Intelligence in healthcare. In S. Mishra, H. K. Tripathy, P. Mallick, & K. Shaalan (Eds.), *Augmented intelligence in healthcare: A pragmatic and integrated analysis* (pp. 475–502). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1076-0_25

Mukherjee, A., Ghosh, S., Behere, A., Ghosh, S. K., & Buyya, R. (2020). Internet of Health Things (IoHT) for personalized health care using integrated edge-fog-cloud network. *Journal of Ambient Intelligence and Humanized Computing, 12*(1), 943-959. https://doi.org/10.1007/s12652-020-02113-9

Ordinance MCTI No. 4,979, of July 13, 2021. (2021). Amends Ordinance No. 4,617, of April 6, 2021, which establishes the Brazilian Artificial Intelligence Strategy and its thematic pillars. Official Gazette of the Union. https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligencia artificial/ebia-portaria_mcti_4-979_2021_anexo1.pdf

Resolution No. 659, of July 26, 2021. (2021). Provides for the National Policy on Health Information and Informatics (PNIIS). https://bvsms.saude.gov.br/bvs/saudelegis/cns/2022/res0659_15_06_2022.html

Sindhosp. (2022). *Proposta Saúde São Paulo 2022*: rumo ao acesso sustentável. https://sindhosp.org.br/linha-do-tempo/#:~:text=O%20que%2%C3%A9%3F,com%20mais%20equidade%20e%20sustentabilidade

Shaik, T., Tao, X., Higgins, N., Gururajan, R., Li, Y., Zhou, X., & Acharya, U. R. (2022). Fedstack: Personalized activity monitoring using stacked federated learning. *Knowledge-Based Systems, 257(*12), 109929. https://doi.org/10.1016/j.knosys.2022.109929

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using Artificial Intelligence: Current state, applications, and challenges. *WIREs Data Mining and Knowledge Discovery, 13*(2), e1485. https://doi.org/10.1002/widm.1485

Tandel, S., Godbole, P., Malgaonkar, M., Gaikwad, R., & Padaya, R. (2022). An improved health monitoring system using IoT. *Proceedings of the International Conference on Innovations and Research in Technology and Engineering, 7*, 1-5. https://doi.org/10.2139/ssrn.4109039

Weenk, M., Bredie, S. J., Koeneman, M., Hesselink, G., van Goor, H., & van de Belt, T. H. (2020). Continuous monitoring of vital signs in the general ward using wearable devices: Randomized controlled trial. *Journal of Medical Internet Research, 22*(6), 1-11. https://www.jmir.org/2020/6/e15471/

World Health Organization. (2021). *Ethics and governance of Artificial Intelligence for health: WHO Guidance*. https://iris.who.int/handle/10665/341996

World Health Organization. (2024). *Exploring the digital health landscape in the WHO European Region: Digital health country profiles*. https://iris.who.int/handle/10665/376540

Wu, Q., Chen, X., Zhou, Z., & Zhang, J. (2022). FedHome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing, 21*(8), 2818-2832. https://doi.org/10.1109/tmc.2020.3045266

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion, 77*, 29-52. https://doi.org/10.1016/j.inffus.2021.07.016

Zheng, X., Shah, S. B. H., Ren, X., Li, F., Nawaf, L., Chakraborty, C., & Fayaz, M. (2021). Mobile edge computing enabled efficient communication based on federated learning in internet of medical things. *Wireless Communications and Mobile Computing*, 1-10. https://doi.org/10.1155/2021/4410894

unesco
Centre
under the auspices
of UNESCO

cetic.br
Regional Center for
Studies on the
Development of the
Information Society

nic.br
Brazilian Network
Information Center

cgi.br
Brazilian Internet
Steering Committee