ISSN 2281-5147

TO USE OR NOT TO USE? RE-USING HEALTH DATA IN AI DEVELOPMENT

Fatma Sümeyra Doğan*

Abstract

This study examines the re-use of health data in the context of AI development, focusing on regulatory frameworks governing this practice under the European Health Data Space. It explores how transparency and the protection of personal data are balanced with the need for innovation in healthcare. By analysing real-world examples and the application of General Data Protection Regulation principles, particularly transparency, this study assesses whether health data can be re-used for AI-driven healthcare advancements without undermining individuals' data protection rights.

Table of Contents

TO USE OR NOT TO USE? RE-USING HEALTH DATA IN AI DEVELOPMENT 177

Abstract 177

Keywords 178

1. Introduction 178

2. Motivations for Heath Data Re-use 179

2.1 Real World Examples of Data Re-use 180

* PhD Researcher and Marie Skłodowska-Curie Action's Fellow at Jagiellonian University under the Legality Attentive Data Scientists (LeADS) project. Her PhD research focuses on health data governance in the EU in relation to emerging technologies. Email: av.sumeyradogan@gmail.com

This work is supported by the European Union's funded project Legality Attentive Data Scientists (LeADS) under Grant Agreement no. 956562.

ISSN 2281-5147

3. Rules and Regulations to Process the Health Data 182
3.1 The General Data Protection Regulation 182
3.2 A New Regulation to Govern the Health Data 183
3.3 Secondary Use under the European Health Data Space 183
4. Data transparency challenges for Health Data re-use 185
4.1 Transparency under General Data Protection Regulation 186
4.2 Transparency vs. Secondary Use of European Health Data Space 187
4.3 Guiding principles for transparency 188
5. Conclusion 189
6. Bibliography 190

Keywords

Secondary Use of Health data – Medical AI – European Health Data Space – GDPR – Transparency

1. Introduction

There is no doubt that healthcare systems could benefit from technological developments to address staff shortages and the increasing number of people who need treatment. AI in healthcare can make a big difference by improving access to treatments and providing more personalized care. It can help doctors diagnose illnesses more accurately, discover new medicines, predict diseases, and obtain support medical professionals by analysing complex data, suggesting treatment options. However, using AI also brings up important issues like protecting people's data and ensuring privacy.

Maintaining the protection of our personal data and privacy carries utmost importance especially in healthcare domain. One of the reasons for this is that we cannot change almost all the data related to our health. For example, our genetic

ISSN 2281-5147

information is fixed from birth and cannot be altered. Similarly, our medical history, such as past surgeries, chronic illnesses and other health conditions we've experienced, is permanent. Unlike passwords or other personal identifiers that can be changed if compromised, these aspects of our health are unchangeable. This makes it crucial to protect this data from misuse or unauthorized access because any breach could have long-lasting or permanent consequences, potentially affecting our privacy, insurance, employment and even the quality of care we receive.

However, at the same time, we must use data to train AI technologies. For example, to develop AI systems that can accurately diagnose diseases, we need access to a vast amount of medical records and imaging data. This data helps the AI learn to identify patterns and make accurate predictions. Similarly, for AI to assist in drug discovery, it must analyse extensive datasets about how different substances interact with the body. Thus, if we want to develop, use and benefit from AI technologies, we must allow the use of data to train these new technologies. To achieve these advancements, the concept of re-using data becomes crucial.

With the increasing potential of AI technologies to in healthcare, the use of health data has become a central concern. While AI can enhance diagnostic accuracy and personalize treatment, its development relies heavily on access to vast amounts of health data. This raises significant legal questions, particularly concerning the re-use of such data beyond its original purpose. In light of the European Health Data Space proposal, which aims to enable the secondary use of health data, this study investigates whether this framework adheres to the General Data Protection Regulation's transparency requirements. Specifically, it examines how transparency can be maintained when data is anonymized or pseudonymized and considers the challenges posed by AI's opacity in data processing.

2. Motivations for Heath Data Re-use

Re-using data refers to the practice of using existing data for purposes beyond its original purposes. For instance, patient records from routine medical visits can be re-used to train AI systems to predict potential outbreaks of contagious diseases such as COVID-19. Additionally, big health datasets can be used to assess post-marketing adverse events and thus the safety of pharmaceutical products. By analysing this data,

ISSN 2281-5147

AI can identify risk factors and help implement preventive measures. For example, AI algorithms can detect patterns in patient records, such as the emergence of symptoms in specific demographics or regions, which may indicate the early stages of an outbreak. The AI can then model potential scenarios, allowing healthcare providers to respond proactively by increasing resources in high-risk areas or recommending targeted interventions like vaccination campaigns or public health advisories. This data-driven approach enables more effective prevention and management of health issues before they become widespread. Additionally, data from fitness trackers and wearable devices can be re-used to enhance AI algorithms that promote healthier lifestyles by providing personalized recommendations on exercise and diet. This re-use of data is essential for advancing AI technologies and unlocking their full potential to benefit society. It allows us to maximize the value of existing data while ensuring that new insights and innovations can be achieved without repeatedly collecting the same information.

In this study, the terms 'secondary use' and 're-use' of data are used interchangeably, as both concepts involve utilizing existing data for new purposes to derive additional value and insights.

2.1 Real World Examples of Data Re-use

In order to provide more concrete examples of the re-use of data in the health sector, we will give a few of them in the following. Google's Automated Retinal Disease Assessment harnesses artificial intelligence to aid healthcare practitioners in detecting diabetic retinopathy, a condition where high blood sugar levels damage the blood vessels in the retina, potentially leading to blindness if left untreated. This technology also has the potential for AI algorithms to further assist clinicians in recognizing other medical conditions (*ARDA*, n.d.). Google collaborated with Moorfields Eye Hospital located in the UK to assemble a dataset of eye retina images. Subsequently, Google Health trained an artificial intelligence system capable of predicting the development of a type of eye disease. A study was conducted to evaluate this system against expert clinicians. The findings suggest that Google's AI system can forecast whether an eye may develop the disease within the next six months as accurately as clinicians (*Using AI to Predict Retinal Disease Progression*, 2020). Additionally, Google explored potential clinical uses of this system, showcasing the promise of AI in preventive medical studies. According to Google this technology now has been used widely in India and

ISSN 2281-5147

practitioners reported that it enables them to examine more patients in a day, which is crucial in over-populated areas (*Dostrzeganie Potencjału - Google*, n.d.).

Moreover, Google has improved this technology and developed an innovative method to predict the risk of a heart attack by analysing images of a person's retina. This advancement also utilizes artificial intelligence to scan the eye and identify risk factors for cardiovascular diseases. By examining the retinal blood vessels, the AI system can accurately predict the likelihood of a heart attack, offering a non-invasive and efficient way to assess heart health, although this method is not yet widely used in clinical practice (Poplin et al., 2018). However, its innovative approach could complement traditional methods and potentially become more common as the technology advances and gains broader acceptance. This method, developed in collaboration with various research institutions, highlights the potential of AI in preventive healthcare. By leveraging retinal images, which are relatively easy to obtain, this approach could revolutionize how heart disease risks are assessed, potentially leading to earlier interventions and better health outcomes. The significance of this development lies in its ability to provide a quick, non-invasive diagnostic tool that can be used widely, especially in settings where traditional methods might be impractical.

On a different study, Altsman and his team at Stanford University utilized statistical analysis and data mining techniques to detect patterns in extensive datasets. They developed a "symptomatic footprint" for drugs that could cause diabetes (Yousefi, 2022, p. 4). By partnering with Microsoft Research, they examined user's anonymous Microsoft search engine logs. Through this investigation, they discovered that combining the drugs named Paxil and Pravachol can lead to diabetes. This conclusion was drawn from the observation that patients taking these two drugs together exhibited a notable increase in searches for terms associated with diabetes, such as "fatigue" and "loss of appetite," indicating high blood glucose levels. This data-driven research provided a crucial, life-saving finding that traditional methods might have missed. Altsman asserts that restricting access to data would be detrimental to research, as data is a vital source of inspiration, innovation, and discovery in medicine. He believes that the ability to analyse new data sources offers unprecedented opportunities to identify problematic drugs or drug combinations much earlier than previously possible (Stanford University, 2016).

ISSN 2281-5147

3. Rules and Regulations to Process the Health Data

Data is a key element in training AI technologies, making it crucial to obtain. This raises the question: how can innovators legally access data? To answer this, we must examineregulations, such as those in the European Union, which will be the focus due to the limitations of this study. (Artice 3(1) of the 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 (General Data Protection Regulation), 2016)

3.1 The General Data Protection Regulation

The General Data Protection Regulation is a law from the European Union designed to protect personal information. It ensures that companies handle data with care and respect. Under the General Data Protection Regulation, individuals have the right to know what data companies have about them, to correct any inaccuracies and to request the deletion of their data if it is no longer needed. Companies must obtain consent before using personal health data and are required to keep it safe from data breach and misuse. They must also be transparent about why they are collecting data and how it will be used. Companies that fail to follow these rules can face significant fines. In essence, the General Data Protection Regulation gives individuals control over their personal data and ensures it is protected.

The General Data Protection Regulation includes special rules for protecting health data because this type of information is highly sensitive. Health data includes medical records, genetic information and any details about an individual's physical or mental health. Because of the sensitive nature of this data, General Data Protection Regulation imposes stricter rules to ensure it is handled with the highest level of care. Under the General Data Protection Regulation, as a general rule, companies and organizations must obtain explicit consent from individuals before collecting or using their health data. This means they must clearly explain why the data is needed and how it will be used and individuals must agree to it. Additionally, health data must be kept secure to prevent unauthorized access or misuse. This includes using advanced security measures like encryption. Encryption is the process of converting data into a coded format that can only be accessed by authorized individuals, ensuring that even

ISSN 2281-5147

if the data is intercepted, it cannot be read by unauthorized parties. The stricter rules are in place to protect individuals' privacy and to prevent any potential harm such as identity theft, discrimination or unauthorized use of personal health information, that could come from the misuse of sensitive health information. By mandating stringent measures for handling health data, the General Data Protection Regulation aims to safeguard personal health information and build trust in how it is managed.

3.2 A New Regulation to Govern the Health Data

The strict protection rules imposed by General Data Protection Regulation on health data have inadvertently inhibited the development of AI technologies by restricting access to the data needed to train and improve these systems. To address this challenge, proposal for the European Health Data Space was introduced to improve healthcare quality and continuity across Europe (European Commission, 2022). Another key reason is to accelerate medical research and innovation. With access to larger and more diverse datasets, researchers can conduct more comprehensive studies, leading to quicker and more significant medical advancements. This collaborative approach, supported by the European Health Data Space, will establish clear rules and frameworks that facilitate the secure sharing of health data across borders and between institutions. This will help in developing new treatments, understanding diseases better and improving public health outcomes. The European Health Data Space also aims to empower individuals by giving them more control over their health data. During the discussion phase at the EU Parliament, an opt-out mechanism was proposed, allowing individuals to choose not to participate in the secondary use of their health data, ensuring that their consent remains central to the process. (*European Health Data Space*, n.d.)By making it easier for people to access and manage their medical records, the European Health Data Space aims to promote modernisation in the healthcare systems of the European Union.

3.3 Secondary Use under the European Health Data Space

The secondary use framework under the European Health Data Space is designed to facilitate the re-use of health data for purposes beyond the initial care of patients, while ensuring robust data protection and privacy. Once the proposal fully adopted by the Member States, health data will be collected from various sources, such as hospitals, clinics and wearable devices. Before this data can be used for secondary

ISSN 2281-5147

purposes, it undergoes a process of anonymization or pseudonymization to enhance confidentiality. Anonymization involves removing all personal identifiers so that individuals cannot be identified, while pseudonymization masks identifiers by replacing them with codes (pseudonyms) that can only be re-linked to the original data under strict conditions.

Strict protocols will be established to determine who can access the data and for what purposes. Researchers must apply for access through a regulated process, which is overseen by independent authorities known as health data access bodies. These bodies are established under the governance frameworks of the European Health Data Space, ensuring their operation is guided by strict legal and ethical standards. Composed of experts in data protection and relevant scientific fields, these bodies are independent from research institutions which allows them to impartially evaluate applications. This board assesses the potential benefits of the research or project against privacy risks, as well as other potential risks such as data misuse, ensuring that data is used responsibly. Approved users access the data through secure environments, often referred to as data safes or data access platforms, which implement advanced security measures such as encryption, secure login and monitoring to prevent unauthorized access or data breaches. The European Health Data Space establishes comprehensive governance frameworks that outline the responsibilities of all parties involved in data handling and usage.

The framework promotes transparency by requiring public disclosure of who is using the data, for what purposes and the outcomes of their research or projects. Accountability measures are in place to address any misuse of data, including penalties and corrective actions, such as revoking access to data or imposing fines. By enabling the safe and legal re-use of health data, the European Health Data Space aims to drive innovation in healthcare, support the development of new treatments and technologies and enhance public health strategies. This approach maximizes the value of existing data while maintaining high standards of data protection and privacy, ensuring that the benefits of data re-use are realized without compromising individual rights.

ISSN 2281-5147

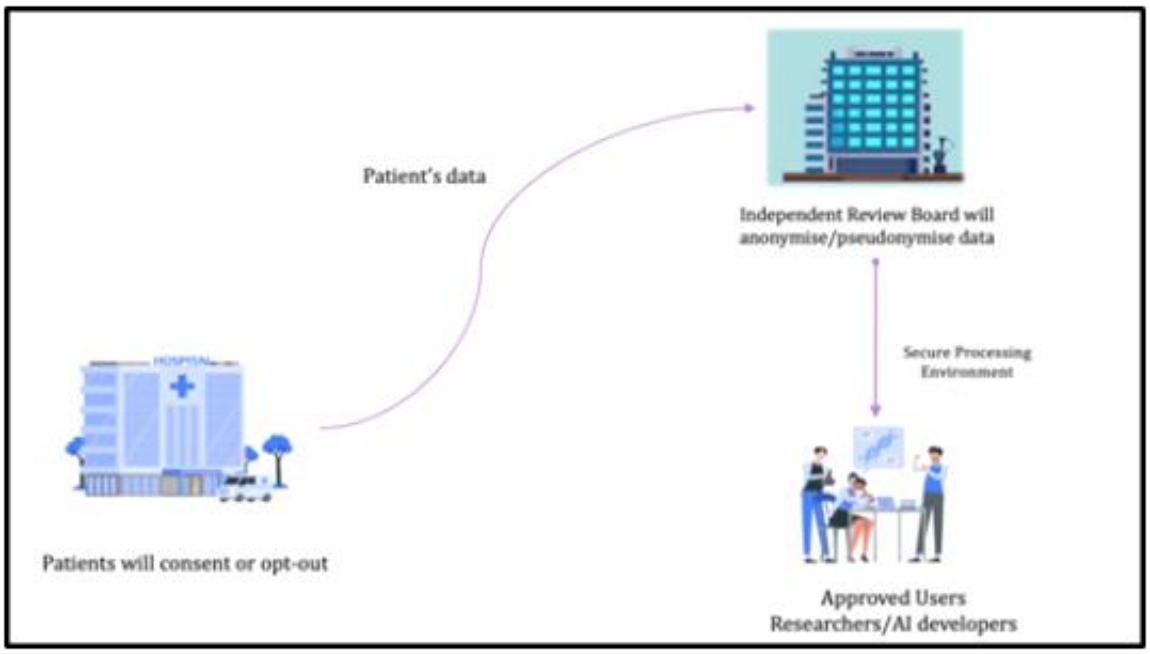

Figure 1- An example for the re-use of health data under the European Health Data Space proposal

4. Data transparency challenges for Health Data re-use

Understanding transparency in data processing is crucial as it forms the backbone of responsible data management practices. Under General Data Protection Regulation, transparency means that data controllers must provide clear, accessible and comprehensive information to individuals about how their personal data is collected, used and shared. This principle is particularly relevant when considering the legal and ethical considerations of the secondary use framework under the European Health Data Space. As the European Health Data Space enables the re-use of health data for research and other secondary purposes, it must comply with stringent transparency requirements mandated by the General Data Protection Regulation. Specifically, Article 14 of the General Data Protection Regulation sets forth detailed obligations for data controllers to inform individuals about the collection and use of their data, when the data is obtained indirectly. This requirement ensures that individuals are fully aware of how their health information is being utilized and protected within the European Health Data Space framework, thus maintaining trust and upholding privacy standards.

ISSN 2281-5147

4.1 Transparency under General Data Protection Regulation

Transparency in data processing is a fundamental principle aimed at ensuring individuals understand how their personal data is collected, used and protected. This principle is essential for building trust between data subjects, who are the individuals whose data is being processed, and data controllers, which are the entities that determine the purposes and means of processing personal data.

Data controllers must communicate clearly and openly about their data processing activities. This involves informing individuals about what data is being collected, the purposes for which it is collected, how it will be used and who will have access to it. Transparency also requires explaining the legal basis for data processing, which can include consent - especially critical for health data- from the data subject, the necessity for the performance of a contract, compliance with a legal obligation, protection of vital interests, public interest or legitimate interests pursued by the data controller.

Additionally, individuals must be informed about their rights regarding their personal data. These rights include access to their data, correction of inaccuracies, deletion of data (known as the right to be forgotten), restriction of processing and data portability, which is the right to receive their personal data in a structured and commonly used format and to transfer that data to another data controller. Transparency also encompasses information about how long personal data will be retained and the security measures in place to protect it. This ensures that individuals are aware of the lifespan of their data and the steps taken to safeguard it against breaches and unauthorized access.

When personal data will be shared with third parties, data controllers must disclose this information, specifying who the third parties are, the purpose of sharing and how the data will be protected in the process. Providing contact information for the data protection officer or another responsible entity is crucial, allowing individuals to seek further information or lodge complaints about data processing activities.

If data processing involves automated decision-making, including profiling - where personal data is used to evaluate certain aspects of an individual, such as their behaviour, preferences or health-, they must be informed about this aspect. They should understand the logic involved, the significance and the potential consequences of such processing. By ensuring these aspects of transparency, data controllers help

ISSN 2281-5147

individuals make informed decisions about their personal data and exercise their rights effectively. Thus, the transparency principle fosters trust, accountability and compliance with General Data Protection Regulation.

4.2 Transparency vs. Secondary Use of European Health Data Space

The balance between transparency and secondary use in data processing is a crucial and often challenging aspect of today's data management practices. As mentioned, transparency requires that data controllers clearly inform individuals about how their data is collected, used and shared. Secondary use, on the other hand, involves using data for purposes beyond the original context in which it was collected, such as for research or analytics, which can be distinct from the initial data collection purpose.

This trade-off arises because achieving high levels of transparency often necessitates detailed disclosures about data use, which can sometimes conflict with the need to manage secondary use. This challenge becomes even more pronounced when data users have profit-driven motives. For instance, consider a company that collects health data from wearable devices for research purposes. Initially, the company informs users that their data will be used to improve health monitoring technology and to conduct general health studies. This initial disclosure is straightforward and focuses on the primary purpose of data collection. However, the company's long-term plan involves using this data to develop targeted marketing strategies for health-related products and services, such as personalized dietary supplements or fitness programs. To maximize profits, the company might not fully disclose these secondary intentions to users at the time of data collection. By keeping these plans less transparent, the company can more easily obtain consent from users who may otherwise be hesitant if they knew their data would be used for targeted marketing. In this context, the opt-out mechanism provided by the European Health Data Space offers individuals the ability to refuse secondary uses of their data.

The overarching question this study seeks to address is: For the sake of innovation, should we give up on protecting our personal data and privacy? The European Health Data Space regulation proposal will provide access to health-related data through its secondary use of data framework. It also includes several safeguards, such as anonymization and pseudonymization. Nonetheless, it remains unclear how the transparency principles of the General Data Protection Regulation will be adhered to

ISSN 2281-5147

within this new framework. This study aims to investigate whether the secondary use schema of the European Health Data Space proposal aligns with General Data Protection Regulation's transparency rules.

As discussed, this goal does not have a straightforward answer. Transparency is a multifaceted concept and ensuring it when data is processed in anonymized or pseudonymized forms presents particular difficulties. For example, it is challenging to maintain transparency about data use when the data itself has been altered to remove personal identifiers. Additionally, AI technologies are known for their opacity in how they process and analyze data, making it difficult to fully understand and communicate how data is being used. Developers benefiting from secondary use of health data sould be more transparent about their purposes for using the data.

4.3 Guiding principles for transparency

To address these transparency challenges, we can draw insights from the U.S. Food and Drug Administration, an authoritative body that regulates medical devices and their software, has established guiding principles for transparency in machine learning-enabled medical devices. These principles highlight the need for clear communication about how devices are used, including their intended purpose, development, performance and the underlying logic of their algorithms. Drawing from the Food and Drug Administration's recommendations, it is clear that effective transparency involves not only providing relevant information about a device's functionality and performance but also ensuring that such information is accessible, timely and comprehensible to users.('Transparency for Machine Learning-Enabled Medical Devices', 2024)

Incorporating these principles into the European Health Data Space framework could address some of the transparency challenges. For example, clear and ongoing communication about how health data is used and how AI systems make decisions could help bridge the gap between data anonymization and user understanding. Policymakers and developers should consider adopting strategies similar to those outlined by the Food and Drug Administration, such as enhancing user interfaces to present information more clearly, providing timely updates, and using human-centered design principles to make data use more transparent.

ISSN 2281-5147

Ultimately, this study seeks to determine whether a balance can be struck between leveraging health data for technological advancements and maintaining stringent transparency and data protection standards as mandated by the General Data Protection Regulation. By aligning with best practices in transparency, as suggested by institutions like the FDA, we can better safeguard individual rights while unlocking the full potential of health data for technological progress.

5. Conclusion

In summary, the European Health Data Space represents a significant step forward in the responsible use and protection of health data. We have explored the balance between transparency and anonymity, highlighting the importance of clear communication and robust anonymization techniques to protect individual privacy while enabling valuable medical research and technological advancements. Key legal and ethical considerations, such as adherence to the General Data Protection Regulation and the transparency requirements of it, ensure that data processing within the European Health Data Space framework is conducted legally.

Looking to the future, the potential of the European Health Data Space to revolutionize healthcare is immense. By facilitating the secure and legal re-use of health data, the European Health Data Space can drive innovations in personalized medicine, improve public health strategies and support the development of new treatments and technologies. However, challenges remain, particularly in maintaining the delicate balance between transparency and privacy and ensuring that data re-use does not compromise individual rights.

As the landscape of health data use continues to evolve, it is crucial for individuals to stay informed about their data rights and the measures in place to protect their privacy. By understanding the principles of transparency and the importance of data protection, we can promote a more informed and engaged public that supports the re-use of health data. This collective awareness will help ensure that the benefits of the European Health Data Space are realized while safeguarding the privacy and data protection rights.

ISSN 2281-5147

6. Bibliography

ARDA: Using Artificial Intelligence in Ophthalmology - Google Health. (n.d.). Retrieved 11 May 2024, from https://health.google/caregivers/arda/

Dostrzeganie potencjału—Google. (n.d.). Retrieved 11 May 2024, from https://about.google/intl/ALL_pl/stories/seeingpotential/

European Commission. (2022, March 5). *Communication from The Commission to The European Parliament and The Council A European Health Data Space: Harnessing the power of health data for people, patients and innovation*. https://doi.org/10.1163/2210-7975_HRD-4679-0058

European Health Data Space: Council and Parliament strike deal. (n.d.). Retrieved 17 April 2024, from https://www.consilium.europa.eu/en/press/press-releases/2024/03/15/european-health-data-space-council-and-parliament-strike-provisional-deal/

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158–164. https://doi.org/10.1038/s41551-018-0195-0

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 (General Data Protection Regulation), EP, CONSIL (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Stanford University. (2016, March 30). *Harnessing big data to better understand what happens when we mix drugs*. Welcome to Bio-X. https://biox.stanford.edu/highlight/harnessing-big-data-better-understand-what-happens-when-we-mix-drugs

Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. (2024). *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles

ISSN 2281-5147

Using AI to predict retinal disease progression. (2020, May 18). Google DeepMind. https://deepmind.google/discover/blog/using-ai-to-predict-retinal-disease-progression/

Yousefi, Y. (2022). Data Sharing as a Debiasing Measure for AI Systems in Healthcare: New Legal Basis. *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 50–58. https://doi.org/10.1145/3560107.3560116

ISSN 2281-5147