eXplainable AI for trustworthy healthcare applications

Doctoral Thesis
by

Cecilia Panigutti

Doctoral Program in Data Science

Supervisor
Dino Pedreschi, Università di Pisa

Supervisor
Giovanni Comandè, Scuola Superiore Sant'Anna

© Cecilia Panigutti, 2022. All rights reserved.

The author hereby grants to Scuola Normale Superiore di Pisa permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

eXplainable AI for trustworthy healthcare applications

by

Cecilia Panigutti

Wednesday 13th April, 2022 11:51

Submitted to the Scuola Normale Superiore di Pisa
in -, in partial fulfillment of the requirements for the
Doctoral Program in Data Science

Abstract

Acknowledging that AI will inevitably become a central element of clinical practice, this thesis investigates the role of eXplainable AI (XAI) techniques in developing trustworthy AI applications in healthcare. The first part of this thesis focuses on the societal, ethical, and legal aspects of the use of AI in healthcare. It first compares the different approaches to AI ethics worldwide and then focuses on the practical implications of the European ethical and legal guidelines for AI applications in healthcare. The second part of the thesis explores how XAI techniques can help meet three key requirements identified in the initial analysis: transparency, auditability, and human oversight. The technical transparency requirement is tackled by enabling explanatory techniques to deal with common healthcare data characteristics and tailor them to the medical field. In this regard, this thesis presents two novel XAI techniques that incrementally reach this goal by first focusing on multi-label predictive algorithms and then tackling sequential data and incorporating domain-specific knowledge in the explanation process. This thesis then analyzes the ability to leverage the developed XAI technique to audit a fictional commercial black-box clinical decision support system (DSS). Finally, the thesis studies AI explanation's ability to effectively enable human oversight by studying the impact of explanations on the decision-making process of healthcare professionals.

Keywords: eXplainable Artificial Intelligence, XAI, Trustworthy AI, Healthcare applications

Publications

1. Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pages 97–110. Springer, 2019

2. Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 629–639, 2020

3. Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657, 2021

4. Cecilia Panigutti, Anna Monreale, Giovanni Comandè, and Dino Pedreschi. Ethical, societal and legal issues in deep learning for healthcare. In *DEEP LEARNING IN BIOLOGY AND MEDICINE*, pages 265–313. World Scientific, 2022

5. Cecilia Panigutti, Andrea Beretta, Fosca Giannotti and Dino Pedreschi. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems (accepted for publication in the Proceedings of 2022 *ACM CHI Conference on Human Factors in Computing Systems*)

Dissemination publications in italian

1. Cecilia Panigutti and Emanuele Bosi. Intelligenza artificiale in ambito diabetologico: prospettive, dalla ricerca di base alle applicazioni cliniche. *Il Diabete Online, Organo ufficiale della Società Italiana di Diabetologia, Medicina traslazionale*, 33(1), 2021

2. Fosca Giannotti, Dino Pedreschi, and Cecilia Panigutti. Ia comprensibile per il supporto alle decisioni: Doctor xai. In *Biopolitica, pandemia e democrazia. Rule of law nella società digitale. Vol. 3: Pandemia e tecnologie. L'impatto su processi, scuola e medicina*, pages 109–120. Il mulino, 2021

Conferences

1. AAAI International Workshop on Health Intelligence, Honolulu, Hawaii, 2019

2. Conference on Data Science and Law, ETH Zurich, Zürich, Switzerland, 2019

3. ACM Conference on Fairness, Accountability, and Transparency Barcelona, Spain, 2020

4. Workshop on Explainable Medical AI: Ethics, Epistemology, and Formal Methods, Leiden, The Netherlands, 2021

5. ACM CHI Conference on Human Factors in Computing Systems, New Orleans, Louisiana, 2022

"*Data science is only as much of a science as it facilitates the interpretation of data – a two-body problem, connecting data to reality. Data alone are hardly a science, regardless how big they get and how skillfully they are manipulated*"

- Judea Pearl

Acknowledgments

Throughout my Ph.D. journey, I have received a great deal of support and assistance.

First and foremost, I would like to thank my supervisors, Dino Pedreschi and Giovanni Comandè. My work would not have been possible without your support, advice, and encouragement. I will be forever grateful for your guidance and wisdom. I feel fortunate to have found supervisors that believed in me and made me feel valued. I would also like to thank Fosca Giannotti, co-PI of the XAI Project, for the support and research opportunities and the members of the panel that helped monitor my research progress: Salvatore Ruggieri, Stan Matwin, Fabio Gadducci, and Anna Monreale.

Special thanks also go to my co-authors. In the chronological order of our first work together: I want to thank Riccardo Guidotti for giving me the initial push I needed to start my research, Anna Monreale for always taking the time out of her busy days to listen to me, and for her unwavering support throughout the past four years, Alan Perotti for his mastery, advice and friendship, Paolo Bajardi and André Panisson, who were present since the very beginning of my professional growth and have always supported me and believed in me. Last but not least, I would like to thank Andrea Beretta for his enthusiasm despite the challenging times and for introducing me to the complicated venture of interdisciplinary research.

Special thanks also go to other members of the Data Science Ph.D. board and researchers connected to the program who have supported my research activities: Francesca Chiaromonte, Piero Marchetti, Lorella Marselli, Emanuele Bosi, Francesca Pratesi, Vittorio Romano, Salvatore Rinzivillo and Daniele Fadda. A special thank goes to Sarah De Nigris, who helped me revise this thesis and provided extremely helpful insights. I also want to thank Mattia Setzu for have always been open to frank exchanges of views and ideas.

I would also like to thank the following people, who supported me personally during this research, and without whom I would not have completed the Ph.D. degree. Firstly I would like to thank my remarkable fellow Data Science Ph.D. candidates with whom I shared the joys and sorrows of Ph.D. life: Jisu Kim, Giorgio

Tripodi, Vasiliki Voukelatou, Gevorg Yeghikyan, Tommaso Radicioni, Luca Insolia, and Elisa Ferrari. I am extremely grateful to have met so many incredible friends in Pisa. To name them all would require far more than a few pages. A special mention goes to Elisa, the cornerstone of my home-away-from-home. I also want to explicitly thank Isabella, Martina, and Giulia for the interesting and intellectually stimulating discussions that enriched my research and worldview. I thank my family, whose profound belief in my abilities gave me the courage to embark on this adventure. Finally, I want to thank Stefano, the companion of my life's adventures, for his love and support throughout the ups and downs of this journey.

Contents

1 Introduction **17**

2 AI transparency and explainability **21**
2.1 Transparent-by-design . . . 22
2.2 Explainable AI . . . 23

3 Thesis Objectives **26**

4 Ethical, legal and societal issues of AI applications in healthcare **28**
4.1 Introduction . . . 28
4.1.1 A practical definition of AI . . . 30
4.1.2 On the importance of ethical considerations when developing AI tools . . . 31
4.2 Main Contribution . . . 32
4.3 AI ethical and legal guidelines around the world . . . 32
4.3.1 US . . . 34
4.3.2 China . . . 36
4.3.3 EU . . . 38
4.4 EU Seven Requirements for trustworthy AI . . . 41
4.4.1 Human agency and oversight . . . 41
4.4.2 Technical robustness and safety . . . 42
4.4.3 Privacy and data governance . . . 43
4.4.4 Transparency . . . 44
4.4.5 Diversity, non-discrimination and fairness . . . 46
4.4.6 Societal and environmental wellbeing . . . 46
4.4.7 Accountability . . . 47
4.5 The AI application lifecycle stages . . . 48
4.5.1 Design stage . . . 49
4.5.2 Development stage . . . 50
4.5.3 Deployment and Maintenance stage . . . 54
4.5.4 Usage stage . . . 55
4.6 Relevant EU legislation . . . 56
4.6.1 Medical Devices in EU . . . 56
4.6.2 Medical Device malfunction . . . 58
4.6.3 Handling health data under the GDPR . . . 60
4.7 Discussion . . . 64

5 A solution to the *black box outcome explanation problem* for healthcare data 67
5.1 Introduction . . . 67
5.1.1 Terminology and definition of the *outcome explanation problem* 68
5.1.2 The local neighborhood . . . 70
5.1.3 Multi-label classification tasks . . . 73
5.1.4 Sequential Data . . . 75
5.1.5 Ontology-linked data . . . 76
5.2 Main contribution . . . 81
5.3 MARLENA: multi-label black box outcome explanation . . . 82
5.3.1 Problem definition . . . 82
5.3.2 Neighborhood Generation . . . 85
5.3.3 Experiments . . . 89
5.3.4 Results . . . 91
5.3.5 Lessons learned . . . 96
5.4 Doctor XAI: sequential and ontology-linked data . . . 99
5.4.1 Ontology use in machine learning and XAI . . . 100
5.4.2 Methods . . . 101
5.4.3 The explanation pipeline . . . 101
5.4.4 Experiments . . . 110
5.4.5 Results . . . 115
5.4.6 Explanation example . . . 118
5.4.7 Lessons learned . . . 121
5.5 Discussion . . . 123

6 XAI to audit clinical decision support systems that are proprietary software 127
6.1 Introduction . . . 127
6.1.1 Bias in healthcare data and algorithms . . . 128
6.1.2 Fairness . . . 130
6.2 Main contribution . . . 131
6.3 FairLens: target user and context . . . 132
6.4 FairLens: pipeline . . . 134
6.5 Use Case: auditing a medical decision support system . . . 139
6.5.1 Dataset: MIMIC-IV . . . 140
6.5.2 Clinical DSS: Doctor AI . . . 140
6.5.3 Local Explainer: DoctorXAI . . . 141
6.5.4 Local-to-global approach: GlocalX . . . 141
6.5.5 Auditing DoctorAI on MIMIC-IV . . . 142
6.6 Validation . . . 149
6.6.1 FairLens Analysis . . . 151
6.7 Discussion . . . 154

7 Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems 156
7.1 Introduction . . . 156
7.1.1 Theoretical models of acceptance and use of technology . . . 157

7.1.2 Trust . . . 159
7.1.3 Trust calibration and AI explanations . . . 160
7.2 Main contribution . . . 161
7.3 Methods . . . 162
7.3.1 Participants . . . 162
7.3.2 Estimation task . . . 162
7.3.3 Experimental design . . . 162
7.3.4 Collected data . . . 163
7.3.5 Interface "Only suggestion"- Dr. AI . . . 165
7.3.6 Interface "Suggestion and explanation" - Dr. XAI . . . 165
7.4 Results . . . 166
7.4.1 Quantitative analysis . . . 166
7.4.2 Open-ended questions insights . . . 169
7.5 Discussion . . . 171

8 Final discussion and future work 173
8.1 Transparency . . . 173
8.2 Auditability . . . 175
8.3 Human oversight . . . 176
8.4 Future work . . . 177

A Appendix A: FairLens use case 2 - auditing a clinical DSS for predicting medical codes from clinical notes 179
A.1 Dataset: MIMIC-III . . . 180
A.2 Clinical DSS: CAML . . . 180
A.3 Local Explainer: The Attention Mechanism . . . 180
A.4 Local-to-global approach . . . 181
A.5 Auditing CAML on MIMIC-III . . . 181
A.5.1 Assessing the DSS performance on the healthcare facility data 182
A.5.2 Identifying problematic groups of patients . . . 182
A.5.3 Identifying systematic sources of error in the selected subgroup 183
A.5.4 Obtaining explanations for systematic misclassifications . . . 183

B Appendix B: user study additional information 185
B.1 Information sheet . . . 185
B.1.1 What is the project's purpose? . . . 185
B.1.2 Procedures . . . 186
B.1.3 Study information . . . 188
B.2 Informed consent . . . 190
B.3 Demographics . . . 193
B.4 Need for cognition (NFC) . . . 194
B.5 Instructions . . . 196
B.6 Tutorial clinical history . . . 199
B.7 Tutorial Interface A: only suggestion of the AI algorithm . . . 201
B.8 Tutorial Interface B: suggestion of the AI algorithm and explanation . 202
B.9 Estimation task . . . 204
B.9.1 Initial estimate . . . 205

B.9.2 Final estimate: only suggestion . . . 206
B.9.3 Final estimate: suggestion and explanation . . . 207
B.10 Trust scale . . . 208
B.11 UTAUT - pt1 . . . 209
B.12 UTAUT - pt2 . . . 210
B.13 UTAUT - pt3 . . . 211
B.14 Explanation satisfaction scale . . . 212
B.15 Open questions . . . 213
B.15.1 Additional open questions for interface B "Suggestion and explanation" . . . 213
B.16 Final open questions . . . 213

Bibliography **214**

List of Figures

3-1 Overview of the thesis structure and relationships between different chapters 27

4-1 Interrelationship of the seven requirements from the EU commission website 40

5-1 Toy example in 2 dimension of a complex decision boundary for a binary classification task (from [302]). The red cross is the instance to be explained, the dotted line is the local surrogate model (linear in this case). 70

5-2 A representation of a branch of the tree-shaped ICD-9 hierarchical ontology: the root is a general condition *Disease* while its children and grandchildren are increasingly more specific conditions. 78

5-3 *(1st plot)* 2D feature space with multi-label instances having 3 types of different labels, the arrow points out the instance to explain. *(2nd plot)* MARLENA-m selects αk neighbors in the feature space. *(3rd plot)* MARLENA-m selects $(1-\alpha)k$ neighbors in the latent space. *(4th plot)* The resulting *mixed neighborhood* obtained merging merge the previous sets of instances. 86

5-4 *(1st plot)* 2D feature space with multi-label instances having 3 types of different labels, the arrow points out the instance to explain. *(2nd plot)* MARLENA-u selects k neighbors using the distance function which combines distances in the features and in the label space generating in one single step the *unified neighborhood*. *(3rd plot)* MARLENA-u perturb the core of k neighbors generating a dense *synthetic neighborhood* around the instance to explain. 87

5-5 Hit and *r-fidelity* varying α for *yeast* and *woman*, upper and lower figure respectively. 92

5-6 Distributions of mean mixed distance among core real neighborhood points. 94

5-7 The explanation pipeline 102

5-8 *(1st plot)* The node corresponding to the randomly selected ICD-9 code (276.4) of the patient is highlighted in red in the ICD-9 ontology graph representation. *(2nd plot)*. The ontological superconcept of the selected ICD-9 is selected and highlighted (276). *(3rd plot)* All ICD-9 codes all having as parent the identified superconcept are selected and removed from the patient (codes 276.1, 276.2 and 276.4). 107

5-9 Example of temporal encoding for a patient . . . 108
5-10 Example of temporal decoding for a patient . . . 109
5-11 ICD-9 ontology. The red dots represent codes occurring in the MIMIC dataset, the orange ones their parent nodes. . . . 112
5-12 Fidelity distribution for the ontological pipeline with different k, perturbation type, and training/test set. . . . 116
5-13 Fidelity distribution for the non-ontological pipeline at different values of k and training set. . . . 117
5-14 Explanation complexity for the ontological pipelines . . . 118
5-15 Explanation example . . . 121

6-1 FairLens as a tool for auditing a clinical decision support system before its deployment in a healthcare facility. Our contribution, in blue, provides to the auditor an instrument to detect and explain systematic ML model biases on protected groups. . . . 132
6-2 Fairlens pipeline: a tool to support human experts investigating if a black-box clinical DSS behaves differently on groups based on protected attributes, highlighting which health conditions are more often misclassified and why. . . . 135
6-3 Distributions of demographic attributes in the auditing data . . . 143
6-4 Normalized disparity scores vs. group sizes with bootstrap outliers bands capturing 50% (light grey) and 95% (dark grey) of the random variability for that group size. The median of the bootstrap distribution is shown as a solid grey line. . . . 144
6-5 Aggregated visualization of the relevant ICD-9 codes for the underdiagnosis of *Heart valve disorders* in over-65 Asian patients. . . . 147
6-6 Sampling procedure and distribution of demographic variables in the training sets and test set. We first extract from the whole MIMIC-IV a test set with 20% of data points. From the remaining points, we extract two training sets with different sampling procedures (sampling A and sampling B). Sampling A produces a training set with artificially injected bias. Sampling B produces a training set with random sampling that respects the same distribution of demographic variables as the original dataset. . . . 150
6-7 FairLens scatterplots for the baseline DSS (a) and biased DSS (b) . . 152
6-8 Average metrics across insurance groups . . . 153

7-1 Technology Acceptance Model (TAM) . . . 157
7-2 Unified Theory of Acceptance and Use of Technology Model (UTAUT) 158
7-3 Flowchart of the estimation task for the two interfaces: only suggestion (blue path) and suggestion and explanation (yellow path) . . . 163
7-4 Static visualization of the *only suggestion* AI interface . . . 165
7-5 Static visualization of the *suggestion and explanation* AI interface . . 166
7-6 Boxplot comparing the WOA (a) the confidence shift after the advice (b) the behavioural intention of use and (c) the explicit trust in the two systems (d). . . . 167

A-1 Normalized disparity scores vs. group sizes with 50% and 95% bootstrap outliers bands. 182
A-2 Aggregated visualization of the relevant terms for the over-diagnosis of *428.0: Congestive heart failure* in Medicare patients between 45 and 65 years. 184

List of Tables

5.1 Real health-related dataset information and black box performance (F1-measure). 90
5.2 Fidelity (mean ± stddev) of *MARLENA-m* and *MARLENA-u* on all datasets. 92
5.3 *s-fidelity* (mean ± stddev) of *MARLENA mixed* and *union* for each dataset. 92
5.4 *r-fidelity* (mean ± stddev) of *MARLENA mixed* and *union* for each dataset. 93
5.5 Hit performance comparison (mean and standard deviation). 95
5.6 Hit performance comparison (mean and standard deviation). 95
5.7 Mean rule length and standard deviation comparison between MARLENA-m and GDT. 95
5.8 Mean rule length and standard deviation comparison between MARLENA-u and GDT. 96
5.9 MIMIC-III characteristics for patients with more than one visit . . . 111
5.10 *Doctor AI* performance on different datasets. 113
5.11 Mean values of fidelity 119
5.12 Mean values of hit 119

6.1 MIMIC-IV: Data from patients with at least two hospital admissions . 140
6.2 clinical DSS performance 142
6.3 Groups with the highest disparity score in each group size bin. All disparity scores marked with $*$ are above the 95th percentile of random variability for the group size. 145
6.4 Groups ranked by disparity scores and most over/under-diagnosed conditions when auditing the black-box 146
6.5 Set of rules produced by DoctorXAI and aggregated by GlocalX to explain why the CCS code 96: *Heart valve disorders* was under diagnosed for over-65 Asian patients by the model DoctorAI. Each row group is a rule with a set of premises, each premise is in the form of ICD-9 $\gtrless$ *threshold_value*. For the human-readable description of each ICD-9 code the reader can consult `http://www.icd9data.com/`. 147
6.6 Performance of clinical DSS trained on the biased and on the baseline training sets 150
6.7 Groups with the highest disparity score in each group size bin for the biased DSS. All disparity scores marked with $*$ are above the 95th percentile of random variability for the group size. 151

6.8 The ranking performed by FairLens using disparity score for the baseline and biased DSS. 151

7.1 Comparison of UTAUT variables for the two interfaces. Median, paired sample Wilcoxon signed-rank test statistics and p-value. 168

A.1 CAML performance on auditing data 182
A.2 Groups ranked by normalized disparity scores for different group sizes and most over/under-diagnosed conditions when auditing the black-box 183

Chapter 1

Introduction

The last decade has witnessed an increasing digitalization of every aspect of our life, including our health. Many new paths to generating and gathering large volumes of health-related data were introduced thanks to the digital transcription of our clinical history into Electronic Health Records (EHR) [178, 177], the widespread availability of smartphones and health apps [23, 24], and the creation of cheap wearable sensors able to track nearly any kind of physiological signal [226, 292, 290]. In addition, the cost of generating reliable omics data (such as genomics, proteomics, and transcriptomics) has dropped [381], allowing the creation of patient-specific multi-omics profiles and opening the door to personalized medicine. The richness of information contained in such data allows gaining unprecedented insight into health and disease conditions. However, the vast amount of heterogeneous data that a single patient generates makes it impossible for any human being to process all that information alone. It is apparent that, in the future, doctors will increasingly need to rely on the help of advanced algorithms able to process and make sense of big volumes of health data. The development of such algorithms is one of the main focuses of Artificial Intelligence (AI) and, in particular, of Machine Learning (ML) and Deep Learning (DL). Recent advances in AI have shown the ability of DL models to successfully solve narrow tasks such as interpreting medical scans [222, 253, 69, 350], pathology slides [34, 84, 57], skin lesions [107, 149], retinal images [147, 9] and electrocardiograms [233, 401]. While some claims have been made about a future where AI will replace doctors [203], AI is more likely to become an essential tool in

doctors' service, allowing them to outsource mundane tasks to algorithms and focus on more serious matters [353, 328]. AI and human doctors will have complementary roles reflecting their strengths and weaknesses. While an AI algorithm can be trained to have a better-than-human vision [285, 250] and will never make a diagnostic mistake stemming from fatigue [399], it will never be able to go beyond the patterns it has learned from already existing evidence and come up with a new solution to a one-of-a-kind clinical case. Most importantly, an AI algorithm will never understand a patient's experience of their illness and establish a human connection, which is at the center of medicine [386, 94]. It is therefore of pivotal importance to develop an AI technology able to work synergistically with doctors.

Current AI technologies have many shortcomings that hinder their adoption in the real world. From a technical point of view, these models can suffer from various issues such as sensitivity to adversarial attacks [118, 294, 232], overfitting to training data, and inability to manage data distribution shifts due to a lack of causal reasoning [62]. Furthermore, the great amount of data needed to properly train and test AI models requires researchers to access big volumes of protected personal data, raising privacy concerns and creating information and power asymmetries between *big tech* companies and the public [209, 373]. AI models might also be sensitive to biased datasets and algorithms, creating fairness issues. It has indeed been repeatedly shown that these models underperform on underrepresented groups of patients and can also learn to perpetuate discriminatory decision-making patterns [344, 259]. Lastly, most of these medical AI systems are generally validated using only retrospective studies on a small number of clinical sites and with a general lack of doctors-in-the-loop [388].

These issues are further exacerbated by the *black-box* nature of most state-of-the-art AI systems. Indeed, these models might have millions of parameters capturing the extreme nonlinearities of the input features, making their internal decision-making process hard to interpret by human beings. The uninterpretability of such models makes it difficult to examine their reliability, identify potential malfunctionings and prevent them from happening again. In the healthcare context, AI-based clinical Decision Support Systems (DSS) having a black-box model at their core pre-

vents the clinician from investigating unexpected findings and perform a differential diagnosis process. Ideally, clinical DSS should enable clinical reasoning and allow for scrutiny in light of the broader clinical context available to the doctor [70].

Acknowledging that AI will inevitably become a central element of clinical practice, this thesis wants to address some of the methodological, ethical, and legal issues related to the design of trustworthy AI applications in healthcare. In particular, we focus on *explainability* as a means of achieving transparency, one of the key requirements for trustworthy AI. The thesis is structured as follows:

In chapter 2 we introduce the reader to the overarching theme of this thesis: AI transparency and explainability in healthcare. While this chapter presents some of the foundational works on AI transparency and explainability, every chapter begins with an overview of the related works specific to the topic being discussed.

In chapter 3, we present an overview of the specific objectives and research questions of this thesis linking them to the following chapters.

In chapter 4, we present how different great world powers like Europe, China, and the United States are tackling ethical, legal and societal issues stemming from the use of AI in healthcare. We then perform an in-depth analysis of the EU ethical and regulatory framework for trustworthy AI. This analysis highlights how considering the socio-technical and legal context is crucial to develop appropriate technical solutions to real-world problems. AI transparency emerges as one of the crucial elements to enable AI to be trustworthy and fair.

Chapter 5 is the technical core of the thesis, where we address the AI transparency problem from a technical point of view. This chapter focuses on enabling *explainable AI (XAI)* techniques to deal with common healthcare data characteristics and tailor them to the medical field. We present two novels XAI techniques that incrementally reach this goal by first focusing on multi-label predictive algorithms and then tackling sequential data and incorporating domain-specific knowledge in the explanation process.

The subsequent chapters broaden the perspective on transparency, bringing together the social, legal and ethical aspects that emerged in chapter 4 and the technical aspects presented in chapter 5.

In chapter 6, we explore the interplay between explainability and fairness by developing a framework that exploits the XAI methods developed in chapter 5 to audit clinical DSS that are proprietary software. Indeed proprietary software impairs transparency in a non-technical way by not allowing the users to inspect the source code of the AI model.

In chapter 7 we investigate the impact of the explanations provided by our XAI methods on the perceived trustworthiness of an AI-based clinical DSS and on the intention of using such an AI system in the medical field. We do this by carrying out an online user study on health professionals following human-computer interaction experimental design principles. Finally, in chapter 8, we discuss the findings of this thesis, and we outline future works.

Chapter 2

AI transparency and explainability

The study of techniques whose goal is to explain (i.e., capability to present in human-understandable terms [98]) the decision-making process of an AI system is as old as the AI field itself [243]. This topic has recently witnessed an increased interest that generated vast literature on AI transparency and explainability. Indeed, the popularity of such techniques matches the increasing use of *black-box* AI systems, i.e., systems whose internal decision-making process is obscure. An AI system might be considered a black box for two reasons:

- The AI system is based on a complex Machine Learning model whose outcomes cannot be understood and interpreted just by looking at its internal parameters. In this case, the lack of transparency reflects a lack of knowledge or understanding of the model's inner knowledge representation.
- The AI system is based on proprietary software. The source code of the model, its specifications and the data used to train it are not available. In this case, the lack of transparency might have nothing to do with the inherent characteristic of the Machine Learning model.

Some Machine Learning researchers have compared the black-box reasoning of AI applications in healthcare to the black-box reasoning of many doctors, claiming that it is impossible to explain all the factors which led a physician to his or her diagnosis [266, 49]. However, being able to explain clinical decisions to patients and be

held accountable for adverse outcomes of their diagnosis are key ethical responsibilities of every doctor. Furthermore, it has been argued that article 22 of the GDPR establishes a right to explanation, making explicability a legal requirement [236]. The need to understand the reasoning behind AI decision support systems, i.e., *explicability*, is also listed as one of the four ethical imperatives of the EU guidelines for trustworthy AI [163].

There are two ways of reaching the level of transparency mandated by the GDPR and suggested by the EU guidelines. The first way is to avoid the use of black-boxes and to use inherently interpretable models instead: the *transparent-by-design approach* [59, 17, 376]; and the second way is applying techniques from the field of *explainable AI* (XAI) [302, 228].

2.1 Transparent-by-design

A model is interpretable if the user can understand and interpret how the inputs are mathematically mapped into the outputs. In the literature, there is a small number of models that are recognized as inherently interpretable: linear models, decision trees, and decision rules [145]. Black-box models such ad Deep Learning models are usually preferred to inherently interpretable ones because they capture highly non-linear relationships between the variables without requiring a feature engineering process and therefore yield higher accuracy with less effort. However, there are some examples of high-performance interpretable models in healthcare. For example, Caruana et al. used generalized additive models with pairwise interactions (GA^2M) to predict pneumonia risk and hospital 30-day readmission generating high-performance interpretable models [59]. Since GA^2M allows visualizing single and pairwise feature interactions with the outcome, the authors were able to identify a dangerous omitted-variable bias present in the real-world data set used to train the algorithm. More specifically, the model classified asthma patients as having a low risk of dying for pneumonia complications because the data set did not contain information on the type of treatment those patients received. This finding highlights the dangers of using black-boxes that do not allow for such exploration

of the model's learned biases in healthcare applications. Another way to achieve a transparent decision-making process is to design the algorithm to mimic the human decision-making process. For example, in [28] the authors trained a Deep Learning model to look at specific aspects of digital mammography images known to be important, based on the physiology of how lesions develop within the breast tissue. These aspects were then used in a case-based reasoning process presented to the physicians to explain the model classification. Furthermore, in her work, Rudin [306] claims that the accuracy interpretability trade-off is a myth. This claim is based on the fact that Data Science is an iterative process that involves many back-and-forths between problem definition, data analysis, and modeling. During this knowledge discovery process, it becomes easier to find a good data representation that allows simpler models to have the same level of performance as the black-box ones. Another representative example of a transparent-by-design model is Bayesian Rule Lists [212], where the trained model consists of an ordered list of if-then rules that describe the decision-making process of the model. Generally, these transparent-by-design models are based on models recognized as inherently interpretable in the literature: linear models, decision trees, and if-then rules. While this approach to model explanation is always ideal, it is not applicable in all scenarios. Building transparent models with competitive prediction performance is particularly difficult in the case of multi-class and multi-label classification problems [410]. Furthermore, this approach to model explanation can not be applied when the final goal is to audit the decision-making process of proprietary software.

2.2 Explainable AI

EXplainable Artificial Intelligence (XAI) is a term coined in 2017 by DARPA for its homonymous program [148]. XAI is a sub-field of Machine Learning that studies the techniques that explain in human-understandable terms the logic used by a black-box AI model in its decision-making process. These techniques are instrumental if it is impossible to develop a high-performance transparent model for the task at hand or understand the reasoning behind a black-box of the second type (proprietary

software).

In the last few years, there has been a surge in the academic literature related to this field [145, 171] and a complete review of all these methods is outside the scope of this chapter. However, we provide an overview of XAI techniques commonly used for AI applications in the healthcare domain.

One approach to medical XAI is the *model-aware* approach. These XAI methods are often applicable only to specific AI models and require access to internal values of the black-box such as the gradients in the convolutional layers [319] or the *attention scores* [360, 26]. These methods have been used to explain the classification of skin cancer histology images [391], to generate heatmaps for chest CT images [411, 375], and to increase the interpretability of AI models employed in medical longitudinal prediction tasks [190, 125, 329]. However, it has been debated that this kind of explanation might lack consistency and meaning [306], and that attention should not be used as an explanation [176, 384].

Another approach to explaining black-box decisions is through the use of an inherently interpretable proxy model (such as a linear model or a decision tree) able to mimic its local or global behavior [87, 228, 302, 140]. These approaches focus on extracting explanations from a black-box model after training, i.e., *post-hoc*. Several methods falling into the post-hoc category are also *model-agnostic*, i.e., they can be applied to any black-box since they analyze only its input-output behavior [302, 228]. Since the *model-agnostic* approach to model explanation focuses only on its input-output behavior, a plethora of methods have been developed to deal with a variety of data sources (relational [303, 143, 18, 268], text [249], images [319, 142], sequences [271] or several of them [302, 228]), and learning problems (binary and multi-label classification, regression, scoring) allowing the user to choose the best explainer for the task at hand. These models are also often local, which means that the provided explanations are valid only for individual predictions and fail to generalize to the whole model's logic. To overcome this limitation, some new XAI methods have been proposed to generalize the local explanations combining them into a surrogate model able to mimic the black-box logic while being interpretable at the same time [325, 324].

Most of the XAI methods listed above, although often used for medical AI applications, are *application-agnostic* meaning that they can extract an explanation from the black-box AI model regardless of the application domain. While the *model-agnostic* approach to XAI offers high flexibility to the use of these methods, the *application-agnostic* approach implies that the specific user needs are not considered [22]. Indeed, a recent survey has shown that machine learning engineers mainly use explanations of black-box AI models to debug their model in the development phase [39]. Nevertheless, debugging the model is only one of the needs expressed in another recent study that analyzed the demands of transparency of several stakeholders [48].

In this thesis, we explore the XAI line of research that develops XAI methods that are not entirely agnostic and tailor the explanations to the medical field, either by incorporating medical knowledge in the explanation process [73, 407] or focusing on specific healthcare data characteristics and use cases [238]. Furthermore, we investigate the efficacy of our explanations on a group of health care professionals. In the next chapter, we present an overview of our contributions.

Chapter 3

Thesis Objectives

In this thesis, we explore the ability of XAI techniques to meet different requirements for trustworthy AI in the context of healthcare applications. The specific objectives and research questions of this thesis are the following:

1. **Analyze the legal and ethical framework for the development of trustworthy AI systems in the healthcare context** (chapter 4). What are the requirements that need to be met in order to develop trustworthy AI-based applications? How XAI can help meet these requirements?

2. **Propose a solution to the *outcome explanation problem* for healthcare data** (chapter 5). How to enable XAI techniques to deal with multi-label prediction tasks? How to deal with sequential data? How to incorporate medical knowledge into the explanation process?

3. **Test the ability of XAI methods to audit clinical decision support systems based on proprietary software** (chapter 6). How can we audit a black-box clinical DSS in order to detect potential biases on different groups and explain its mislabellings on specific data points?

4. **Understanding the impact of XAI on advice-taking in the healthcare context** (chapter 7). How AI explanations impact users' trust in algorithmic recommendations in the healthcare context? How AI explanations impact users' behavioral intention of using the system in the healthcare context?

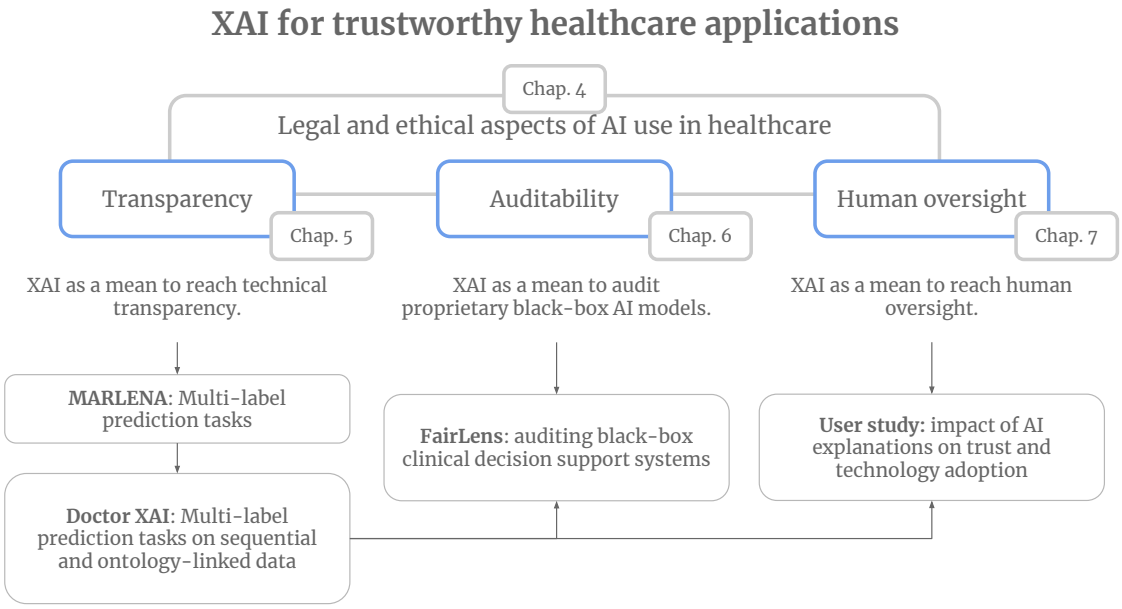


Figure 3-1: Overview of the thesis structure and relationships between different chapters

The analysis of the legal and ethical framework related to the development and use of AI systems in healthcare of chapter 4 embraces and motivates this thesis's other objectives (figure 3-1). Three main requirements related to XAI emerge from this analysis: *transparency*, *auditability*, and *human oversight*. Each chapter builds on the previous one, progressively broadening the view from purely technical to socio-technical and human-centered. Chapter 5 addresses the problem of technical transparency in healthcare, i.e., the use of a complex AI model that is not interpretable. First, the multilabel prediction task is studied. Then, using the acquired insights, DoctorXAI is presented: a local XAI methodology that leverages medical knowledge and deals with multilabel prediction tasks on sequential data. This method is then tested in two scenarios in chapters 6 and 7. In chapter 6, the ability of Doctor XAI combined with a local-to-global approach to audit a fictional commercial black box is explored. Finally, in chapter 7, we study the ability of XAI to enable human oversight through effective explanations. This is done by performing a user study involving healthcare professionals examining the impact of Doctor XAI explanations on trust and intention to adopt a clinical decision support system.

Chapter 4

Ethical, legal and societal issues of AI applications in healthcare

4.1 Introduction

In this chapter, we explore the legal and ethical principles relevant to the development of AI applications in healthcare (objective 1 of chapter 3). Indeed, the implementation of the ethical guidelines, together with the compliance with the legal requirements, can help researchers and developers to design AI systems that can be easily translated into clinical practice [322].

The pervasive use of AI algorithms that exploit and combine sensitive data raises several concerns. For example, the increasing use of e-health apps and wearable devices that directly collect "quasi-health" data [235] (e.g., heart-rate and sleep tracking, breathing regularity, steps count) raises *privacy* issues [215, 197, 370]. When combined with other information such as weight, height, or genetic illness, this kind of data could allow AI algorithms to make inferences about individuals' lifestyles, health conditions, risks of illness, and much more [361, 52, 133, 400]. Furthermore, health data can contain various *biases* due to an imperfect data collection process [280] or to human biases reflected in the data [259]. These biases are difficult to track or discover when fed into opaque and complex AI models, which raises issues of *transparency* and *explainability*. Furthermore, several *liability* issues arise if the AI system is defective, including *medical malpractice* and healthcare providers liability

due to the negligent reliance on AI systems [343, 153, 351].

The importance of considering the ethical and legal implications of the development and use of AI systems is subject to many national and international organizations' recommendations. Some examples are the OECD's *recommendations on the main five values-based principles for the responsible stewardship of trustworthy Artificial Intelligence* [260] signed up by the OECD's 36 member countries, along with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania, and adopted later in 2019, by G20 Trade Ministers and Digital Economy Ministers, the guidelines on regulation of AI [382] released by US administration and finally the ethics guidelines for the development trustworthy AI [163] published in 2019 by the High-Level Expert Group on Artificial Intelligence (AI HLEG) and recently followed by the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence: the AIA (Artificial Intelligence Act) [112].

Since the approach to responsible and ethical AI is not homogeneous across the world, in section 4.3 we analyze the different ideologies, interpretations and guidelines on AI of three of the most influential world powers. In particular, we discuss the European, Chinese and US approaches considering the peculiar societal and ethical issues of AI-based health applications in each of them. This first analysis highlights the main differences between the different interpretations of AI ethical principles worldwide and how these are translated into each legal system.

From this first comparative analysis, the ethical framework delineated by the EU, together with its approach to AI regulation, appears to be the most complete. Therefore, in section 4.4, we focus on the seven requirements for trustworthy AI identified by the AI HLEG. We highlight potential ethical issues of AI applications in healthcare for each requirement and link the AIA and GDPR (General Data Protection Regulation) laws stemming from the analyzed principles.

In section 4.5, we provide a practical perspective by mapping each ethical and legal requirement to the different stages of the lifecycle of an AI product. Finally, in section 4.6, we analyze relevant sectorial legislation such as the Medical Device Regulation and we discuss the impact of the GDPR on AI applications in healthcare.

4.1.1 A practical definition of AI

There is still no universal agreement upon definition of AI. This is because *Artificial Intelligence* is an umbrella term used to indicate a vast family of disparate fast-evolving technologies. The difficulty of pinning down the exact meaning of AI is exemplified by the first attempt to define it for regulatory purposes recently made by the European Commission in its proposal for an AI Act[1]:

> 'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;

Where the current list of technologies in Annex I includes a vast list of approaches ranging from Machine Learning to logic- and knowledge-based and statistical approaches such as Bayesian estimation. The need to resort to an external Annex for a complete picture of the technologies to be regulated is due to the fact that the Commission wants its regulation to stand the test of time and therefore delegates the identification of the relevant technologies to a list that can be easily amended and updated[2]. A more general definition of AI was given by the EU Commission in its communication on *Artificial Intelligence for Europe:* [110]:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).

This definition encompasses all the technologies listed in Annex I of the proposal

[1] TITLE I, Article 3
[2] Article 4: Amendments to Annex I

for an AI Act and therefore, we will adopt it to define what we mean when we use the term AI in this chapter.

4.1.2 On the importance of ethical considerations when developing AI tools

Since no research is performed in a social vacuum, there is no such thing as a *value-neutral* technology. What researchers are interested in studying is an expression of their values and belief system. The technological progress is driven by the cultural, ethical, political and economic interests of society, companies and researchers themselves. Why a certain technology is considered worthy of time and money spending and another one is not? What does a particular technology make it easier to do? There are ethical choices made both in the development phase and in the application phase of any technology. Some choices that might seem harmless and purely technical might not be.

Imagine a researcher is designing a AI model to predict the length of time a patient survives after a liver transplant. The goal of the research is to improve the potential recipient ranking system in order to optimize transplant survival. The choice of the loss function that the AI algorithm must optimize for the task is usually considered a purely technical choice. However, the loss function might implicitly give priority to younger patients, which is an ethical choice, whether right or wrong. Now, imagine that the data set used to train the AI algorithm was collected from a private US-based non-profit organization between 1988 and 1996. The majority of the represented patients are white, furthermore, data shows that the graft survival rate of black patients is significantly lower than those of white ones [252]. The algorithm is highly likely to use this correlation in its optimization phase, even if the race of patients is explicitly removed from the data set [307]. So implicitly, the model will choose white patients over black patients for the transplant even if the observed correlation might have nothing to do with a real causal link between race and graft survival, yet another ethical choice. What is really optimizing the loss function chosen by the developer? The scenario becomes even more complicated if

we consider that the data set is probably outdated since there has been a distribution drift between the last date of data collection (1996) and today [45].

This simple example shows how researchers implicitly embed their beliefs and data biases in the technology they develop. Nevertheless, recent debates have highlighted that many experts on technology feel estranged from the social and policy implications of their work [189]. However, many of the ethical choices listed above are encoded in hard law and therefore are enforceable. Technology development and use are strongly linked to politics and ethical questions on how to advance the good life of individuals or society overall. The field that studies these kinds of ethical questions is *AI ethics*. AI ethics investigates all the ethical questions raised during the development, deployment and use of an AI system. For example, it investigates how values are inscribed into technical artifacts, who is supposed to be held accountable if an AI system fails and the motives behind research goals and findings. The answers to these questions inform regulators around the world.

4.2 Main Contribution

This chapter is based on our work:

- Cecilia Panigutti, Anna Monreale, Giovanni Comandè and Dino Pedreschi. Ethical, societal and legal issues in deep learning for healthcare, 2022 (to appear in the book *Deep Learning in Biology and Medicine*)

4.3 AI ethical and legal guidelines around the world

The ethical and legal guidelines for the responsible use of AI vary across the world. In the past few years, a plethora of private and public stakeholders have produced their reports on the requirements for a trustworthy implementation of AI technologies. Five ethical principles emerge across all the recent literature on this topic: transparency, justice, non-maleficence, responsibility and privacy [185]. However, each country interprets these principles and translates them into its legal system differently [277]. The main dimension along which the different approaches can be

distinguished is the *Regulatory versus innovation* one [362]. For example, the US tends to favor innovation, whereas the EU has a strong regulatory approach. This and many other ethical and regulatory issues arise because of the different socio-economic and political environments in which these technologies are developed. In this section, we analyze the different values systems of three of the most influential world powers that released an AI policy. Before presenting a detailed discussion, we provide a quick overview of the ideologies driving the different approaches to technological innovation in AI:

- **US**: The American approach to AI ethics is influenced by libertarian values that implies minimal regulation of technology from the government. It promotes a "Silicon valley model" which consists in innovating in the regulatory grey zones.

  Conception of human being: Homo Economicus, individualism. "*Move fast, break things first, apologize later*" [20].

- **China**: The Chinese approach to AI ethics is influenced by Confucian values and Chinese socialism ideology. There is a focus on social harmony which implies some elements of moral control and surveillance from the government.

  Conception of human being: collectivist, behaviorist, utilitarian. *"Society as a whole should be mobilized to participate in health affairs, thus contributing to the people's health and the country's overall development"* [408].

- **EU**: The European Union approach to AI ethics is based on the respect of fundamental rights, democracy and the rule of law. In particular, four ethical principles are identified as the most relevant for AI policy: *respect for human autonomy*, *prevention of harm*, *fairness and explicability* [163].

  Conception of human being: Kantian conception of the person as autonomous (freedom, autonomy and dignity).

  "Human dignity is the fundamental concept that provides the framework within which one needs to interpret[...]*European culture and jurisdiction"* [120].

These values and ethical principles guide the implementation of the AI policies and explain the different perspectives on the same issue. It is indeed important to notice that, even though ethical principles are the basis of law, they are not legally binding and therefore are not enforceable. Still, each legal system has many tools to internalize ethical principles in hard law. We will now focus on each of these three regions of the world to discuss their healthcare policies, their approach to AI, the socio-economic peculiarities of each of them and subsequent ethical issues. Furthermore, we give an overview of the most important pieces of legislation concerning AI applications in healthcare for each of them.

4.3.1 US

Many US based big-tech companies such as Google [136] and IBM [170], that develop AI solutions for healthcare, have drafted their ethical policy to address the many concerns about the safe corporate use of healthcare data [209, 61]. Even if their declared goal is to develop healthcare applications which improve the quality of care, reduce healthcare costs and are beneficial for society overall, some potential conflicts of interests are yet to be acknowledged and enforced by legally binding standards [244]. In 2020, the Trump administration released the draft for the *"Guidance for Regulation of Artificial Intelligence Applications"* [382]. The memorandum discourages any *"regulatory or non-regulatory actions that needlessly hamper AI innovation and growth"* coherently with the US approach of prioritizing innovation over regulation.

> Agencies must avoid a precautionary approach that holds AI systems to such an impossibly high standard that society cannot enjoy their benefits. Where AI entails risk, agencies should consider the potential benefits and costs of employing AI, when compared to the systems AI has been designed to complement or replace.

Note that the approach expressly discourages a "precautionary approach" and relies on a traditional cost-benefit analysis. The document list ten principles that federal agencies should consider when they decide how and whether regulate AI

applications. These ten principles are *Public Trust in AI*, *Public Participation*, *Scientific Integrity and Information Quality*, *Risk Assessment and Management*, *Benefits and Costs*, *Flexibility*, *Fairness and Non-Discrimination*, *Disclosure and Transparency*, *Safety and Security* and *Interagency Coordination*. The innovation oriented approach becomes very apparent in the *Risk Assessment and Management* principle:

> It is not necessary to mitigate every foreseeable risk; in fact, a foundational principle of regulatory policy is that all activities involve trade-offs. Instead, a risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits.

These guidelines were following the plan on AI regulation and standards released in February 2019 by the US National Institute of Standards and Technology (NIST) which explicitly addressed the ethical, societal and legal concerns of the development of AI technologies stating that

> While stakeholders in the development of this plan expressed broad agreement that societal and ethical considerations must factor into AI standards, it is not clear how that should be done and whether there is yet sufficient scientific and technical basis to develop those standards provisions.

The regulatory aspects for the AI applications that are classified as medical devices, are regulated by the US Food and Drug Administration (FDA). In January 2020, the FDA released a discussion paper titled *"Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device"* [115]. While the current regulatory framework is being updated and will be defined in the months to come, the ethical concerns are already known. Indeed, in the US the biggest ethical concerns derive from the potential health information asymmetry between companies and individuals and from the systemic biases present in the data sets. An information asymmetry creates a power

asymmetry in favor of big companies like Google [124] that collected several sensitive data on its users' health through their queries on medical conditions. The combination of this knowledge together with other users' interactions on the Internet could potentially infer very personal and sensitive information, blurring the lines that distinguish health and non-health data [235]. This information was legally collected exploiting HIPAA (Health Insurance Portability and Accountability Act) privacy and security rules limited reach to non-traditional healthcare data [78]. The issue is that many users were unaware that their search history would have been processed to infer their health status and have lost the control over this information usage. Potentially, health data could radically change an individual chance of obtaining health insurance, or even employment. These concerns are made all more serious by the fact that currently the US does not provide universal health care to its citizens. Furthermore, health data contain many biases (to be discussed in section 6.1.1) and also reflects the systemic ones. If not properly addressed, the algorithms trained on biased data will learn and perpetrate them in their decision-making process.

4.3.2 China

In late 2016 China released a blueprint of its healthcare strategy *Healthy China 2030* intending to set public health as a priority and to shift the focus from disease treatment to disease prevention [347]. In particular, the blueprint states that the healthcare industry efforts should concentrate on early disease detection, diagnosis and treatment. The blueprint declared goal to prioritize public health is in line with previous efforts of the Chinese Government that quadrupled its healthcare budget between 2009 and 2017 [349]. These initiatives are a response to the increased number of pollution-related illnesses due to poor air quality, concerns about the health management of the country aging population [251] and the inequality of healthcare services access between the rural and urban areas. As declared in the 2017 *New Generation Artificial Intelligence Development Plan (AIDP)*, China plans to use AI as a tool to deal with these health issues. The document, released by the Chinese State Council, outlines China's strategy to become the world leader in the field of AI, among its explicit strategic goals there is the desire to build an *Intelligent Health*

and Elder Care Systems and to design an ethical framework for the use AI to be encoded in hard law [378, 71]. The AIDP guidelines are supposed to be enacted by the private sector [304]. Indeed in the same year of its release, China's Ministry of Science and Technology partnered with the multinational Chinese-based company Tencent to foster AI research in medicine, in particular, to develop computer vision applications for medical diagnosis [184]. The AI-based applications developed in the last years by Tencent are listed in a recent report written for the 2019 special theme on Medical Innovation of the Global Innovation Index [168]. The document reports China's intention to employ AI technologies for triage, clinical decision support system, drug discovery, increasing hospital management and operational efficiency. Furthermore, China plans to manage health knowledge diffusion through the medical information platform *Tencent Medipedia* [348] and to monitor users' health conditions through wearable devices. Aside from the potential dangers of centralizing health information on one platform, the main ethical concerns are related to the subtle distinction between health monitoring and surveilling citizens' health behavior in the name of the common good. Similarly to the Social Credit System initiative [88] fostered by State Council, AI technology could enhance the remote control power of the Chinese Government over citizen behaviors deemed "unhealthy". For example, the lack of physical activity could be monitored using AI applied to inertial sensors data, such as accelerometers and gyroscopes of wearable devices [298]. This danger is relevant for all the so-called *social diseases* (e.g., type II diabetes and obesity) [304]. Another episode that exemplifies the surveillance power of the Chinese Government on its citizens' health behavior is the recent use of AI technologies to fight the Coronavirus epidemic spread. The Government collaborated with Chinese big-tech companies to develop a black box AI system able to classify each citizen according to his risk of being infected. This classification was used to generate a Q.R. code that helped the police to enforce the quarantine [247, 298]. Furthermore, facial recognition technologies paired with contactless temperature detection helped police to identify potential virus carriers which were breaking the law. A similar trade-off between individual rights and social responsibility affects users' right to share their health data only after informed consent.

Indeed, if health data are considered a public good, then they might be collected from unaware users to train sophisticated AI algorithms that could benefit society at large [304, 67, 158].

4.3.3 EU

The European Union is probably the world leader in regulating the ethical principles of AI and in influencing the international discussion on this topic [169]. EU's propensity to code in hard law its ethical principles on AI has raised some concerns about the fact that this regulatory focus might be an obstacle to innovation [318]. However, the European Commission sees the encoding of ethical principles in AI as a competitive advantage that will foster consumers' trust in EU products and as an incentive for companies to create innovative products that satisfy these rules. Indeed, the regulatory approach of the European Commission is thought to foster a trustworthy technology and harmonize its adoption across the Union. The most recent example of hard regulation that also impact AI application is the *General Data Protection Regulation (GDPR)* [2]. The GDPR came into force in May of 2018 and it regulates the processing of personal data in the European Union protecting EU citizen's privacy and stipulating that every EU citizen has the right *not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*. In a similar effort to draft new laws for AI regulation, in June 2018, the European Commission appointed a "High-Level Expert Group on AI" (AI HLEG) to put forward its AI strategy. In the first half of 2019, the group defined seven requirements for trustworthy AI [163], also containing a pilot version of an assessment list for practical use by companies [262]. This document was well received by companies across Europe that contributed to it with their comments and proposals [4, 108]. In the guidelines, three components for the implementation of trustworthy AI are identified: *lawful*, *ethical* and *robust*. The EU approach to ethical and trustworthy AI is fundamental-rights based and human-centered:

> The human-centric approach to AI strives to ensure that human values

> are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoy a unique and inalienable moral status.

This approach promotes research and innovation by putting in place proper safeguards that protect European citizens' rights and freedom. For example, the document explicitly mentions a potential risk of mass surveillance by the Government powered by AI as a critical concern, as opposed to the Chinese approach to surveillance in the name of *social harmony*. It also explicitly mentions the *asymmetries of power or information such as between employers and workers, or between businesses and consumers* as another critical concern. The four ethical principles identified as relevant in the documents are the *principle of respect for human autonomy*, the *principle of prevention of harm*, the *principle of fairness* and the *principle of explicability*. These principles are further operationalized in the white paper *"On Artificial Intelligence - A European approach to excellence and trust"* [111]. In the document, the European Commission outlines its action plan to foster AI use in the framework of European law and ethical values. The plan includes an increased budget for AI research spending of 70 percent. The White Paper specifically addresses the healthcare sector, identifying it as a high-risk sector that needs further legislation refinements. The seven principles identified by the AI HLEG are represented in Figure 4-1.

Using the White Paper as a reference, in April 2021, the European Parliament and the Council released a *"proposal for the regulation laying down harmonised rules on artificial intelligence"* also known as the *Artificial Intelligence Act* (AIA) [112]. The proposal follows a risk-based approach by prohibiting certain practices and heavily regulating others. The stark difference between the EU AIA and the US and Chinese approaches to ethical AI is particularly evident in two key points. First, under the AIA, AI-based social scoring systems such as the Chinese Social Credit System initiative are explicitly prohibited as contravening Union values. Second,

Figure 4-1: Interrelationship of the seven requirements from the EU commission website

the AIA prescribes the *identification and analysis of the known and foreseeable risks associated with each high-risk AI system*, in contrast with the US innovation-first approach.

In the next section, we will analyze the seven principles identified by the AI HLEG as foundational for trustworthy AI highlighting their impact on AI applications in healthcare. We take the European approach to trustworthy AI as a reference because we consider the ethical framework outlined by the AI HLEG the most complete. Furthermore, the effects of EU digital regulations based on such framework usually transcend its confines [5]. The high impact of such regulations around the world is exemplified by the GDPR, which came into full effect in May 2018 and rapidly became a world standard [216, 257]. The process of externalizing EU laws outside its borders is also known as the *Brussel effect* [47]. This process is due to the fact that EU digital regulations also apply to AI providers established outside of the EU providing services to users in the EU.

4.4 EU Seven Requirements for trustworthy AI

The foundations that lay the seven requirements of the EU approach to trustworthy AI are the fundamental rights prescribed by the *Charter of Fundamental Rights of the European Union* [109]. We will now go into the details of each requirement and we will highlight potential AI scenarios impacted by each of them relevant to the healthcare sector.

4.4.1 Human agency and oversight

This requirement reflects the *principle of respect for human autonomy* protecting the fundamental rights of EU citizens and laying the foundations for all the other requirements. It is based on three sub-requirements:

- **Fundamental rights** AI applications should respect fundamental rights. This requires that during the design phase of the application the developers carry out a *fundamental rights impact assessment*, including the protection of personal data and the right to have these data *processed fairly for specified purposes.*

- **Human agency** AI applications should promote human autonomy. This requirement is directly linked to the fundamental right of freedom of the individual which implies that *human beings should remain free to make life decisions for themselves* [163]. This means that AI applications in healthcare should be designed as part of a decision support system allowing end users to make informed autonomous decisions. This principle is also reflected in the article 22 of the GDPR on *"Automated individual decision-making, including profiling"* [80] stating:

  > The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

- **Human oversight** AI applications should allow the human user to have control over the process. This implies that proper human safeguards should be put in place to prevent unintended adverse effects of the AI system. This requirement is in line with the human-centered design promoted in the guidelines and with the right to obtain human interventions in cases ruled by article 22 of the GDPR.

4.4.2 Technical robustness and safety

This requirement is perhaps the most relevant for AI applications developers. It asks them to operationalize the *prevention of harm* principle by paying attention to four key aspects of technical robustness and safety:

- **Resilience to attack and security** AI system developers should prevent system hacking and adversarial attacks. Three targets of attacks are identified: the data (data poisoning), the model (model leakage and flaws) and the underlying infrastructure (hardware and software). In article 15, the AIA prescribes high-risk AI systems to be resilient to *attempts by unauthorised third parties to alter their use or performance by exploiting the system vulnerabilities*. This is a relevant scenario to healthcare applications since medical imaging has been show to be susceptible to targeted adversarial attacks [30, 219].

- **Fallback plan and general safety** AI system developers should put in place a proper fallback plan to cope with adversarial attacks and unexpected situations. This implies an assessment of potential risks (accidental or malicious use of the technology) and a plan to manage the situation. For example, AI could request human intervention before proceeding. The fallback plan should be tested and proper measures for effective redress in case of adverse outcome should be put in place.

- **Accuracy** AI systems should have a high accuracy and should report whenever its outcome/prediction is inaccurate. In the context of the guidelines, the term “accuracy" does not refer to the standard metric used to evaluate Machine

Learning models, it refers to the system ability to perform accurate decisions. This means that the proper definition of system accuracy depends on the task the application is performing.

- **Reliability and reproducibility** AI systems should be both reliable and reproducible. In the context of the guidelines, AI systems are considered *reliable* if they work properly given a specific set of conditions and are considered *reproducible* if under the same conditions they consistently provide the same outcome. The AI system reliability and reproducibility should be constantly monitored and tested, if there are scenarios where the AI system does not meet the standards, such conditions should be reported.

4.4.3 Privacy and data governance

This requirement is in line with Article 7 and 8 of the *EU Charter of fundamental rights* [109] on the *"Respect for private and family life"* and the *"Protection of personal data"*, which are a reflection of the principle of prevention of harm applied to privacy. Data protection is also regulated by the GDPR, along with other directives, across all EU. The guidelines further prescribe special care for sensitive data (some of them include religious, sexual and political orientation, age and gender) that might be inferred from users' digital traces and used to discriminate them. In order to be compliant to this requirement two key aspects should be considered:

- **Privacy and data protection** Since health data are considered among the most sensitive ones, it is of paramount importance to ensure privacy and data protection throughout the entire lifecycle of the AI, eventually performing a data protection impact assessment. We further go into the details on how to deal with health data in Section 4.6.3.

- **Data governance: access, quality and integrity of data** Data governance is the process of managing the data used by an organization. This includes putting in place protocols for *data access* (who can have access to the data), *data quality* (data free of bias, absence of mistakes in the data) and

data integrity (compromised data, data hacking) assessment. Article 10 of the AIA prescribes that the training, validation and test set used in the development of high-risk AI systems should take into account the specific setting of use. Specifically the geographical, behavioural and functional setting. The geographical setting might be relevant to certain AI health applications such as dermatology [11, 192]. Indeed, the presentation of dermatologic diseases in darker skin types might be very different than those in lighter skin types. However, if properly assessed, it has been shown that some AI-based health applications can be safely used across very different populations. For example, in [36], the authors trained an AI system for the automated detection of diabetic retinopathy in retinal images. The model was trained on data from patients from Singapore, while the clinical validation study was performed on diabetic patients from Zambia. The algorithm showed promising generalization results even though these two populations differ in country income status, screening programs set in place and race.

4.4.4 Transparency

This requirement reflects the *principle of explicability* of the EU guidelines for trustworthy AI. Transparency should be applied to every stage of the AI lifecycle, indeed it prescribes the possibility to have a complete view on the whole system. In order to be compliant to this requirement the following aspects should be considered:

- **Traceability** All the steps required to implement an AI application should be properly documented. In the context of a AI application for healthcare, this includes documenting the data collection process and how it was labeled, the choice of the AI architecture together with the optimization algorithm used and how the data was split in order to train, validate and test the model. In the eventuality that the model's wrong outcomes negatively impact a patient's health, it is necessary to understand the reasons behind that decision. In this case, it might be useful to keep track of the model's history of decisions to trace back to a common origin of the mistakes. This is aspects is highlighted by

article 11 and 12 of the AIA, which prescribe to keep a technical documentation up-to-date and to perform automatic recording of events through logging.

- **Explainability** Two levels of explainability are identified: the first one refers to the technical ability to understand the AI decision-making process, while the second one refers to the ability to explain how the human decision-maker interacts with the AI decision-support-system and how (s)he is influenced by it. Their combination contributes to the global transparency of the business model employed. The guidelines prescribe to pay special attention to the AI applications that have a high impact on human lives, for example, in the healthcare context. This aspect is clearly illustrated in article 13 and 14 of the AIA. Indeed, article 13 prescribes that and high-risk AI system should be designed to allow the user to appropriately interpret its output. Furthermore, article 14 prescribes the design of appropriate human-computer interfaces to allow human oversight.

- **Communication** It should always be explicit when a user is interacting with an AI system, and the user should always have the option to opt-out. Most importantly the AI limitations and actual capabilities must be appropriately communicated to avoid overconfidence and overreliance on the AI, which can affect both patients and healthcare professionals. This is considered important for both high-risk AI systems and for AI systems that interact with a natural person, as detailed in article 52 of the AIA. Consider, for example, *Babylon Health*[3], a personalized health-care service that provides an AI-powered chat-bot that operates a triage of patients through guided questions and redirect them to real physicians or pharmacists. It can also give health-care advice, e.g., *how to deal with common cold*. However, the user is aware that the initial interaction is with an AI system and (s)he always has the possibility to further continue the conversation with a real physician.

[3]https://www.babylonhealth.com

4.4.5 Diversity, non-discrimination and fairness

This requirement is in line with the *principle of fairness* listed in the EU guidelines and with article 21 of the EU Charter of Fundamental Rights on Non-discrimination that states that:

> Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

The prevention of discrimination entails three main aspects:

- **Avoidance of unfair bias** Ideally, AI applications in healthcare could facilitate access to better healthcare services increasing societal fairness. However, since these applications hugely rely on the quality of the data they were trained on, they could provide unfair and biased outcomes.
- **Accessibility and universal design** The AI applications should be designed to include the widest possible range of individuals following *Universal Design* principles [338] by taking into account people with diverse abilities, skills, age and size.
- **Stakeholder participation** All the relevant stakeholders affected by AI applications should be involved in their design and maintenance.

4.4.6 Societal and environmental wellbeing

This requirement addresses the environmental costs and the societal risks related to AI applications by extending the principles of fairness and prevention of harm to the broader society. It has been estimated that the carbon footprint of training NLP Deep Learning models is equivalent to the one of a trans-American flight [340]. This has important consequences the environment and consequently on people health. Three important aspects must be considered to be compliant with this requirement:

- **Sustainable and environmentally friendly AI** In the design stage of an AI system there should be an environmental impact assessment (resource usage, energy consumption, carbon footprint). According to a recent study, AI could help realize many *Sustainable Development Goals* [371], for example it could enable the 3rd goal of ensuring *healthy lives and promote well-being for all at all ages* through early detection of diseases, treatment personalization and increasing the quality and accessibility of essential health-care services.

- **Social impact** If an AI system directly interacts with humans a social impact assessment needs to be performed. The end-user needs to know that (s)he is interacting with an AI system and the limits of that interaction. In healthcare, a good doctor-patient relationship is crucial to reduce disease-related anxiety, especially for life-threatening diseases [113]. This requires an interaction with a human doctor. The way the AI is embedded in the clinical setting may be the difference between an increased quality of care and a devastating patient experience. For example, if the AI application interacts directly with the patient giving him or her the diagnosis without an explanation and without the proper communication could hurt the patient's mental health.

- **Society and democracy** AI could potentially increase economic and democratic inequalities. For example, although AI applications in healthcare could enable an easier access to basic healthcare services in rural areas [36], but this could also exacerbate the differences between patients that can afford human care and patients that can not afford it.

4.4.7 Accountability

The accountability requirement prescribes that appropriate mechanisms to identify the responsibility for AI systems' outcomes are put in place during their whole lifecycle. In particular it outlines three sub-requirements:

- **Auditability** It should be possible to assess the algorithms, the data and design processes. This is linked to the previous requirement of transparency and it is a necessary step to ensure the ability to redress.

- **Minimisation and reporting of negative impacts** It should be guaranteed the ability to safely report negative outcomes of an AI decision, eventually fostering an algorithm impact assessment proportionate to the risks posed by the AI.

- **Trade-offs** In the design stage of the AI system it is necessary to acknowledge all the possible trade-offs between the previously listed requirements. This sub-requirement catalyzes, without mentioning it expressly, the only precautionary consideration of the guidelines of the AI HLEG, since they clearly ban the development, deployment and use of an AI system in forms that do not have an ethically acceptable trade-off. An ethically unacceptable trade-off, for instance, is one that undermines the "the essence of the fundamental rights and freedoms" or is not "a necessary and proportionate measure in a democratic society" (see article 23 of the GDPR).

- **Redress** The possibility of adequate redress in case of adverse or unfair outcomes should be guaranteed. For example, in accordance with the requirement on diversity, non-discrimination and fairness, if the AI application fails to address the discrimination bias present in the training data set, the individual should be enabled to ask for effective redress against the machine decision.

4.5 The AI application lifecycle stages

Each stage of the journey from prototype to real-world clinical application of the AI system requires to pay attention to different ethical and legal issues. In the *EU requirements for trustworthy AI*, the AI HLEG identified three stakeholders that should play a role in the guidelines implementation: *developers* (researchers and software engineers), *deployers* (any organization that use AI in their products) and *end-users* [163]. To each of these stakeholders correspond one or more stages of the lifecycle of the AI product. The main four stages for a AI application are the following:

1. **Design**: in this stage the goals of the application are identified, data is ac-

quired, the architecture of the AI model is chosen.

2. **Development**: in this stage the AI model is developed validated and tested. In an academic setting this might be the final stage before writing a research paper.

3. **Deployment and Maintenance**: in this stage the AI model is embedded in an application and becomes a product placed on the market. The product needs to be monitored and updated if needed.

4. **Usage**: in this stage the final product reaches the end-user.

4.5.1 Design stage

In the *design stage* it is crucial to involve all the relevant stakeholders. For a AI application in healthcare whose final goal is to be deployed on the market, an ideal team would include Machine Learning engineers, domain experts such as clinicians and medical researchers, hospital administrators, experts of the legal domain (for regulatory advice) and the future end-users of the application. An interdisciplinary team is also essential to identify relevant clinical scenarios and to prevent possible data analysis pitfalls. For example, developing a model that learns to associate end-of-life treatments to a high risk of mortality is not useful in a real-world clinical scenario since the care team already has this information [385]. In this stage, it is also important to consider the ethical implications of the application. It might be useful to go through the *trustworthy AI assessment list* set up by the European Commission [163] to make sure that all the ethical requirements are satisfied up-front. In particular, Machine Learning engineers should focus on the *prevention of harm* principle and put in place proper safeguards in case of unintended adverse outcomes and malicious use of the technology they are designing. Since health data is considered one of the most sensitive personal data, special attention should be paid to potential privacy risks. More in general, the EU Commission prescribes a *X-by-design* (privacy-by-design, security-by-design, ethics-by-design) approach for AI applications. In other words, in this stage, Machine Learning engineers should

consider both the system's functional requirements and its ethical and legal requirements. The data collection also takes place in the design stage. As previously mentioned, data might contain all sorts of biases. To prevent discriminatory or unintended adverse outcomes, the EU Commission envisages the following requirements for data collection:

- Ensure that the AI application is trained on a sufficiently broad and representative data set.
- Ensure privacy and personal data protection performing a privacy risk assessment on the data.
- Keep record of how and why the data was selected.

4.5.2 Development stage

The goal of the development stage is to develop, validate and test the model. In this stage, the developers need to implement the strategies defined in the previous design step in order to develop a trustworthy Machine Learning model. First of all, *transparency* must be guaranteed during the whole developing process. This is crucial to ensure *traceability* and the correct allocation of liability. To this end, it is important to create appropriate technical documentation and to share data and source code of the AI application (accordingly with proprietary rights). This good practice is optimal to also guarantee *reliability* and *reproducibility* of results. The documentation should contain all relevant information as prescribed by article 11 of the AIA.

Ensure reproducibility To ensure the reproducibility of results it would be optimal to test the model performance against those of benchmarks state-of-the-art models. In this regard, there has been a recent effort of the Machine Learning community to develop such benchmarks models on freely accessible data [188, 201, 173, 374] for many healthcare applications [155, 296, 264]. This aspect is also very important in case the development stage is the last stage before writing an academic paper: if possible, the data and the source code used in the experiments should be shared

with the scientific community. We are aware that this is not possible in most of the cases, especially for healthcare applications that perform their experiments on real-world healthcare data that contains sensitive information. However, a recent good practice to solve this issue is emerging in academic works that develop AI applications for healthcare: the performance of model is reported both on private data sets and on freely accessible data sets, then the pre-processing routines to run the source code and the source code itself are publicly released. The documentation needed to reproduce the results should also include the random seeds as well as the hardware used in the training phase [33].

Proper evaluation of the model In the evaluation phase of the model, the developers should carefully choose the appropriate evaluation metric. Article 15 of the AIA prescribes to declare the levels of accuracy and the relevant metrics in the instruction for use of high-risk AI systems. This metric should take into account if the data set is unbalanced and if it reflects clinically relevant measures. Consider, for example, a AI application that classifies patients' chest x-rays images as having or not having lung cancer. The choice of the appropriate metric changes if the clinical setting is screening or confirmatory. In a screening setting, i.e., a setting where a large number of asymptomatic people are being tested for potential disease, it might be preferable to have a higher sensitivity in order not to miss a person at risk. However, in a confirmatory setting, i.e., a setting where an individual is being tested for a definite diagnosis, it might be preferable to have high specificity [273]. In any case, the AUC score, one of the most used metrics in Machine Learning, does not provide any relevant information in any of the two clinical settings. Furthermore, it is important to prevent *label leakage* when splitting the data into training, validation and test set. Many healthcare application tasks require a patient-level split instead of a random observation-level split. For example, if the AI model must be trained to identify the disease in a chest x-ray image, developers should take into account that one patient may have contributed with more than one image to the data set and thus, a patient-level split is needed [385]. Finally, when the model performance is reported, it is important to also specify the context in which the model was trained

and validated (e.g. single-center data set, adult vs pediatric population, etc.) or in other words the clinical cohort used to develop the model.

Ensure traceability and liability As suggested by the EU commission and prescribed by the AIA for high-risk applications, in order to track back the origin of a potential malfunctioning and to guarantee the determination of liability, the providers of an AI system should:

- Document the training methodologies as well as the testing and validation techniques (article 10 of the AIA on data and data governance).
- Ensure clear information on the application limits and capabilities, for example information about the system robustness to adversarial attacks and about the reproducibility of its results (article 15 of the AIA on accuracy, robustness and cybersecurity).
- Report the goal of the application and the conditions under which it is expected to function as intended (article 13, par. 3 of the AIA).
- Report all the relevant metrics employed in the development of the application (article 15, par 2. of the AIA).

Ensure transparency The transparency requirement is strongly connected to the *explainability* requirement, (the need to understand the reasoning behind AI decision support systems) that is fundamental when Machine Learning models are opaque and incomprehensible to humans, i.e., they are *black boxes*. As a consequence, during the development of AI models it becomes mandatory to take into consideration this aspect by implementing techniques that help in providing tools for explaining the model behavior or the reason of the model decision [145, 79]. In particular, article 13 of the AIA prescribes:

> High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.

This article is linked to the provision of article 14, that states:

> High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.

Such human-machine interface tools will need to allow the human operator to understand the AI output, therefore they will need to incorporate an appropriate explainability technique tailored to the end-user needs [314]. However, transparency does not only means *explicability*. Indeed, we recall that the two reasons for which an AI system might be considered a black box are the following:

- The AI system is based on a complex Machine Learning model whose outcomes cannot be understood and interpreted just by looking at its internal parameters. In this case, the lack of transparency reflects a lack of knowledge or understanding of the model's inner knowledge representation.
- The AI system is based on proprietary software. The source code of the model, its specifications and the data used to train it are not available. In this case, the lack of transparency might have nothing to do with the inherent characteristic of the Machine Learning model.

The transparency requirements prescribed by articles 13 and 14 of the AIA are still valid for the second type of black box AI system. This means that the high-risk AI system users should have enough information to allow the correct interpretation of the system behaviour. However, the intellectual property rights are protected by article 70 of the AIA.

Ensure privacy and fairness In this stage, it is also necessary to address the *privacy* and *fairness* issues, identified during the *design stage*. To this end, developers should implement protection techniques to mitigate privacy and unfairness risks. Some of theses techniques operate data transformations to eliminate risks from training data while others mitigate the risks changing the learning process of

the AI model. However, since these techniques can lead to a degradation of the model accuracy, before applying any mitigation strategy, developers should first assess the possible risks of privacy leakage or unfair behavior in order to focus their intervention only where necessary [286, 310]. The combination of risk assessment and mitigation strategies provides the ingredients to define and develop AI based systems with guarantees of compliance with existing legislation and ethical frameworks.

4.5.3 Deployment and Maintenance stage

The goal of the deployment stage is to put the AI model on the market. In this stage, it becomes fundamental to consider all the relevant pieces of legislation that we discuss in Section 4.6. For high-risk AI systems, the AIA explicitly states that:

> High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity in accordance with the generally acknowledged state of the art. The level of accuracy and accuracy metrics should be communicated to the users.

According to article 9 of the AIA, the AI system provider should establish a post-market monitoring system consisting of

> a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating.

Furthermore, some new forms of evaluation of the model might be necessary to prove the clinical utility of the final tool. For example, a recent work argues that a AI model that has just been tested using the training-validation-test set split lack proof of clinical validity [273]. This claim is mainly due to two reasons: first, because the data set might contain all sorts of biases that prevent the model from generalizing well in clinical practices, and second, because such an application should prove to be useful to the patients' health outcomes. In particular, the authors suggest an evaluation of the entire patient treatment strategy involving the AI application through randomized controlled trials.

As prescribed by the AIA, once deployed, the AI application should be monitored to allow for system maintenance. In 2016, the World Health Organization (WHO) released its first guidelines on *Monitoring and Evaluating Digital Health Interventions* [263]. Even if the focus of these guidelines is on digital health intervention at a national level, the considerations they set out are also relevant to smaller-scale applications (up to the penultimate stage of the application maturity). In particular, the WHO guidelines states that, as the digital health application matures over time, also the monitoring activity needs to evolve. Indeed, in the context of a AI application, the monitoring activity has to verify that such application is working as intended by continuously assess if there is a degradation of the system performance over time. At the same time, the monitoring activity should also take into account the system's compliance with the ethical requirements. In other words, *deployers* should put in place auditing processes able to assess both system's technical performance and system's ethical and legal requirements compliance. Indeed, some features of the data set describing a certain population might undergo a significant data distribution drift over time. This means that the relationship between the inputs and outputs of the model changed due to external factors, e.g., the relationship between patients' features and their probability of survival to a renal transplant might change because of medical innovations. Such kind of drift might affect both medical and ethical aspects. Therefore, they might lead to the degradation of the AI model performance (e.g., accuracy degradation [60]) and degradation of the ethical risks mitigation, such as privacy protection degradation or fairness degradation. This monitoring is essential because, based on the result of this assessment, the *deployers* might need to retrain the model.

4.5.4 Usage stage

According to the AIA, users of high-risk AI systems are required to use such systems according to their intended use and following the documentation provided. The users has also responsible for monitoring the AI system according to instructions of use (article 29 of the AIA on obligations of users of high-risk AI systems). Indeed, AI users should be informed and aware about possible ethical and legal risks derived

from the use of that application. For example, user should have the opportunity to know if data used to make a prediction are stored and/or used to update the learning model by continuous learning techniques [272]. Finally, when the AI application directly interact with natural persons, they should be aware that it is not an interaction with a human being.

4.6 Relevant EU legislation

In its White Paper on Artificial Intelligence [111], the EU Commission sets out its policy objectives regarding regulation. This policy has been implemented in the AIA [112], the act which regulates high-risk AI systems in the EU. High-risk applications are identified with two criteria: the sector of application and if the application is employed in a way that significant risk is likely to arise. The focus on high-risk AI systems is particularly relevant for AI applications in healthcare since these are explicitly mentioned as such. However, the AIA is not the only relevant regulation for AI systems in healthcare. Indeed, every software intended to be used in a medical device needs to be compliant with the *Medical Device Regulation* [3], and personal healthcare data needs to be handled under the *GDPR* [2]. While we highlighted some of the AIA requirements in the last section, we now focus on these other relevant regulations.

4.6.1 Medical Devices in EU

Under the EU law, software intended to be used in a medical device needs to fulfill some requirements to be compliant with the *EU Medical Device Regulation* (MDR). A first innovation MDR entails, is that all software intended to be used for medical purposes by the manufacturer is considered a medical device under its regime. This means that also AI applications for healthcare are considered medical devices and are subject to the same regulation. In particular, the software applications, that fall under the definition of a medical device, are all the applications developed for:

- Diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease

- Diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability
- Investigation, replacement or modification of the anatomy or of a physiological or pathological process or state
- Provide information by in vitro examination of specimens derived from the human body, including organ, blood and tissue donations

However, the regulation specifies that software intended *for general purposes, even when used in a healthcare setting or software intended for life-style and well-being purposes, is not a medical device.* A clear example is a wearable device that tracks vital signs (e.g., blood pressure, heartbeat, oxygen saturation) and offers advice on lifestyle and sleep habits, as opposed to a device performing exactly the same tasks and functions and developed the same way but that is intended for the above mentioned medical uses. The latter is subject to MDR legal and safety rules, whereas the first one is not (even if they are created out of the same dataset and methods)[4]. The Annex I of MDR "General safety and performance requirements" requires that such software

> shall be designed to ensure repeatability, reliability and performance in line with their intended use. In the event of a single fault condition, appropriate means shall be adopted to eliminate or reduce as far as possible consequent risks or impairment of performance.

Furthermore, the regulation requires a document containing detailed information regarding *test design, complete test or study protocols, methods of data analysis, in addition to data summaries and test conclusions.* In particular, the regulation mentions the need for information regarding stability, performance and safety. These requirements guarantee accountability in the medical device monitoring systems. Also, according to this information and characteristics, the MDR classifies medical devices into different risk classes (the higher the class, the higher the risk it entails):

[4]Note that the MDR applies also to the provision of a diagnostic or therapeutic service offered by Information Society (IT) services as defined in point (b) of article 1(1) of Directive (EU) 2015/1535 or by other means of communication (article 6 MDR). Thus a Machine Learning-based software offering a diagnostic or therapeutic service is subject to the same regulation.

- **Class I** - *Low risk*
- **Class IIa** - *Low to medium risk*
- **Class IIb** - *Medium to high risk*
- **Class III** - *High risk*

In general, software is classified in Class I. However, software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes and software intended to monitor physiological processes are classified in Class IIa. Finally, if such decisions have an impact that may cause respectively serious deterioration of a person's state of health/surgical intervention and death or an irreversible deterioration of a person's state of health they are classified medium to high (Class IIb) or high risk devices (Class III). If the nature of variations of the monitored vital physiological parameters is such that it could result in immediate danger to the patient, it is classified as Class IIb.

It then becomes apparent that it is essential to consider the intended use of the developed medical device together with the information regarding its stability, performance and safety. These are indeed the bases of the risk-benefit analysis needed to obtain and maintain marketability and to eventually investigate liability for defective medical devices. This risk-benefit analysis is also vital to secure sufficient financial coverage for the eventual malfunctioning of the medical device, which should be proportionate to its risk class, type of device and size of the enterprise.

4.6.2 Medical Device malfunction

The legal definition of a medical device malfunction must be sought in the coordination of the two definitions given by the MDR and by the *Product Liability Directive*[5]. The MDR defines *device deficiency*[6] in terms of:

> Any inadequacy in the identity, quality, durability, reliability, safety or performance of an investigational device, including malfunction, use errors or inadequacy in information supplied by the manufacturer.

[5] article 6, Council Directive 85/374/EEC
[6] article 2 n. 59

while according to the Product Liability Directive:

> A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account, including: (a) the presentation of the product; (b) the use to which it could reasonably be expected that the product would be put; (c) the time when the product was put into circulation.

In case of a malfunction, current EU legislation on liability for defective products[7] states that:

> If a defective product causes any physical damage to consumers or their property, the producer has to provide compensation irrespectively of whether there is negligence or fault on their part.

Therefore, the producer of the medical device is the subject that should be held accountable if the consumer is harmed.

However, in case of medical devices that use AI it might be difficult to identify the origin of the defect and hold accountable the device manufacturer or the developer, if they are different [79]. Note that, in principle, the manufacturer of a complex device, that incorporates more components, is usually considered the solely liable entity for a defective product. Finally, note that the safety level requested is high for medical devices[8]. Accordingly, stability, performance and safety information are particularly relevant for developers since the reference to inadequacy of performance widens the notion of defective device.

Even if the health-related AI application does not fall under the definition of medical device, some other relevant laws still apply. Consider for example the case of a AI application trained on "unrepresentative health data" [64], i.e. data containing features that describe only a particular ethnic group and that does not generalize well outside of that ethic group. The *EU Race equality directive*[9] might apply to this kind of application if their scope remains in the protected domains. The fact that training

[7] Directive 85/374/EEC
[8] EUCJ Joined Cases C-503/13 and C-504/13, of 5 March 2015
[9] Directive 2000/43/EC

datasets should be sufficiently representative is also both directly mentioned in the White Paper [111] and prescribed in the AIA. Another health-related example of this kind is a AI model which targets the wrong dietary advice [25] to a consumer that has a chronic pathology such as diabetes or has a mental illness such as Anorexia nervosa, the *Unfair Commercial Practices Directive*[10] and the *Consumer Rights Directive*[11] might apply. Similarly, the Products Liability Directive applies, although the notion of defectiveness can have a smaller scope for non-medical device health-related products. Note, again, that less regulated Machine Learning-based products might remain on a higher slippery slope for possible unethical uses.

4.6.3 Handling health data under the GDPR

Whenever an AI system processes EU citizens' personal data, such a system is regulated by the GDPR (General Data Protection Regulation) [2]. Therefore it is of pivotal importance to understand what type of data is considered *personal data* under the GDPR and what constitutes *processing* of such data. Furthermore, according to the GDPR, even in a research context, it is essential to determine a valid *legal basis* for the lawful processing of personal data. We will now go into the details of these issues focusing on AI-based health applications in a research context.

Data protection as a fundamental right In the EU, the protection of a natural person personal data is considered a fundamental right by article 8 of the Charter of Fundamental Rights of the European Union [1]:

> Protection of personal data - data should be processed fairly and for specified purposes and on the basis of consent or some other lawful basis.

The fundamental right to personal data protection is also linked to the right to privacy (article 7 of the Charter). Indeed having technical tools for data protection is needed in order to ensure data privacy, i.e. to guarantee that only authorized users can access sensitive and personal data.

[10] Directive 2005/29/EC
[11] Directive 2011/83/EC

Personal data and data processing under the GDPR As previously mentioned, the notions of privacy and data protection in the GDPR are directly connected to those of *personal data* and *processing*[12] which are expanded with respect to previous EU rules. Indeed, the GDPR applies only to personal data, which is any *information relating to an identified or identifiable natural person*. The GDPR increases the spectrum of data considered as personal data in its definition of *identifiable natural person*:

> an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Compared to previous definitions, this one notably adds *location data*, and *genetic identity* to the list of identifiers of a natural person. The GDPR defines data processing as:

> any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

Consequently, almost all forms of AI-based processing of personal data fall within the scope of the right to data protection, regardless of whether the right to privacy is impaired. If the such processing is carried out for research purposes, it enjoys some simplifications. From now on, we will focus on this latter case.

Legal bases for personal data processing under the GDPR It is worth noticing that the GDPR does not apply to *anonymous data*. Recital 26 offers a reidentifiability test:

[12] article 4 par. 1 and 2

> To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

In case data is not anonymous, the GDPR requires a **legal basis** for any data processing, i.e., it is necessary to identify the scenario where data processing is legally permitted. We will now identify the different possible legal bases (article 6 and 9 GDPR) for the AI-based processing of personal data for research purposes in the medical field.

- **Consent** The consent of the data subject, although sometimes problematic, has always been a very important legal basis also for research. It is *any freely given, specific, informed and unambiguous indication of the data subject's wishes*[13] by which data subjects accept the processing of their personal data. Consent must be given for one or more specific purposes[14]. This general requirement of the GDPR is problematic for data-intensive activities and for data reuse also within Machine Learning. However, recital 33 of the GDPR contributes to the greater flexibility of consent in the context of scientific research:

  > Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.

  The practical consequences of this recital appears to be the legitimacy of broad consensus formulas possibly covering reuse provided that they cover specific areas of research and the relevant ethical standards are respected. Nonetheless,

[13] article 4, par. 11
[14] article 6, par. 1(a)

consent is not always an appropriate legal basis because it is always withdrawable[15] and, in the absence of another legitimate legal basis, any further processing after such withdrawal would be unlawful, requiring the immediate erasure of personal data. Note that data deletion following the withdrawal of consent affects the model developed, the training implemented, the technical choices made. Thus, the whole development and deployment stages of the lifecycle of the AI application need to be reviewed accordingly, taking into consideration that also the assessment of privacy, bias and fairness should be influenced.

- **Special Categories of Personal Data** Given the issues of consent, other legal bases for data processing for AI-based applications in the health domain can be more suitable. While the processing of special categories of personal data such as genetic data and data concerning health is prohibited by article 9, par. 2(h) and (i) allow the use such data in health research without consent when the law provides for an exception respecting the essence of the right to data protection.

- **Public Interest and Legitimate Interests** Article 9 of the GDPR establishes that Union or Member States law may provide alternatives legal bases such as the public interest[16] and the legitimate interests of the holder[17]. However, to rely on the public interest there must be a legal basis found in Member States or EU laws.

Further processing

A central issue for AI systems is that of legitimate further processing and therefore the reuse of personal data. Concerning this aspect, we need to distinguish two scenarios:

- **Research Context** Article 5.1(b) explicitly states that further processing of personal data for scientific or historical research or statistical purposes is

[15] article 7, par. 3
[16] article 6, par. 1(e)
[17] article 6, par. 1(f)

compatible with the initial purposes if aligned with Article 89(1) that relaxes some constraints. Further elaboration, a presumption of non-incompatibility for research is therefore allowed. To benefit from it, the safeguards set out in Article 89 and Recital 156 must be respected, including the demonstration that it was not possible to use anonymous data.

- **Non-Research Context** In case the further processing in the Machine Learning processing is not for research purposes, there is no presumption of compatibility with the original processing. The developer must positively evaluate (in a demonstrable way for accountability purposes) the provisions of Article 6.4.

Data governace under the GDPR The GDPR also imposes the establishment of a real governance structure for personal data: obligation to demonstrate compliance[18], hypothesis of appointment of a data protection officer[19], rules on data breaches[20] and sanctions, including fines of up to 20 million euros or 4% of total turnover[21].

4.7 Discussion

This chapter explored the ethical principles and legal requirements relevant to the development of AI applications in healthcare. The analysis brought to light some important challenges and issues. First of all, the approach to AI ethics worldwide is deeply heterogeneous. In the era of globalization, this could be an obstacle to the regulatory harmonization of such technologies that can, in turn, result in a fragmented global market, undermining trust in AI-based systems and slowing down their adoption in real-world clinical scenarios. Often, such different approaches to AI ethics depend on cultural differences, which can also have an impact on the definition of some values, such as *fairness*, which is one of the most difficult ethical values to

[18] articles 5, 13 and 30
[19] articles 37-39
[20] articles 33 and 34
[21] article 83

be uniquely defined across the world. Furthermore, our analysis highlighted that the human component is, at the same time, too present in the biases often discovered in the data and too far removed from the decision-making process.

Considering the EU guidelines for trustworthy AI and the recent proposal for an AI act, the analysis highlighted three requirements particularly relevant to XAI research.

First of all, the *transparency requirement* and the related *explainability sub-requirement*. In this context, XAI methods might help achieve the technical ability necessary to understand AI decision-making. However, several research gaps need to be addressed to use these methods to interpret AI-based healthcare applications correctly. One of these gaps is purely technical and concerns the applicability of existing XAI techniques to typical healthcare data (we further explore this issue in chapter 5). However, solving technical issues does not guarantee appropriate XAI solutions. Indeed, as highlighted by our analysis, the transparency requirement is also strongly linked to the *human oversight* and *fairness requirements*.

We explore the relationship between XAI and fairness in chapter 6. In particular, we study how XAI tools can be used to audit a black-box AI system based on proprietary software in a healthcare setting to prevent possible fairness issues. This effort is also motivated by the *accountability requirement* and the related *auditability sub-requirement*. This issue shifts the focus from an algorithm-centered perspective on XAI to a sociotechnical and human-centered one.

Furthermore, our analysis highlighted that the EU guidelines explainability requirement also implies the ability to explain how the human decision-maker interacts with the AI decision-support system and how (s)he is influenced by it. This aspect is connected to the *human oversight requirement* and is also clearly illustrated in articles 13 and 14 of the AIA. Indeed, we recall that article 13 prescribes that and high-risk AI system should be designed to allow the user to interpret its output appropriately, and article 14 prescribes the design of appropriate human-computer interfaces to allow human oversight.

This requirement again reveals the limits of a purely technical approach to XAI. An interdisciplinary approach is fundamental to study explanations effectiveness in

enabling human oversight. Indeed, the goodness of an explanation does not lie in the XAI method but in the perceptions of the person receiving the explanation. We explore this topic in chapter 7, where we adopt a human-computer interaction point of view to test the impact of AI explanations on trust and intention to adopt the technology in the context of clinical decision support systems.

Chapter 5

A solution to the *black box outcome explanation problem* for healthcare data

5.1 Introduction

As mentioned in the previous chapter, we begin our journey to test the ability of XAI techniques to meet the requirements for trustworthy AI in healthcare with a technical take on the problem. In this chapter, we study how to solve the *black box outcome explanation problem* (defined below) for healthcare data (objective 2 of chapter 3). Indeed, the healthcare domain poses many unique challenges that require novel XAI techniques to be addressed. Furthermore, such a high-stake domain requires a dedicated effort to develop tailored solutions to perform a sanity check of black-box models beyond mere performance [138, 352]. Since most of the successful applications of AI are in the domain of image processing and computer vision [149, 9], most of the XAI techniques for healthcare data have been focusing in the medical image domain [332, 301]. However, few of them can be directly applied in the medical domain [21, 103] and even fewer focus on other types of healthcare data. For example, healthcare data often presents peculiar features such as sequentiality (section 5.1.4), multi-label predictions (section 5.1.3), and links to structured background knowledge (section 5.1.5).

5.1.1 Terminology and definition of the *outcome explanation problem*

In this section, we introduce the terminology we use to identify the main components of the *outcome explanation problem*, i.e., the problem of providing an explanation for a specific black box outcome. Names, symbols and definitions follow those introduced in [144]. Consider a statistical learning problem [157] where a *ML algorithm* is used to train a *ML model* from data (i.e., the ML algorithm already used the available training data to optimize its internal parameters with respect to its error function). As explained in greater detail in chapter 2, we call *black box model* or *predictor* an already trained ML model whose decision-making process is obscure. More specifically, given an input space $\mathcal{X}^{(m)}$ (the set of all possible inputs $x \in \mathcal{X}^{(m)}$ of the model) and an output space $\mathcal{Y}$ (the set of all possible outputs $y \in \mathcal{Y}$), the black box model is a function $b : \mathcal{X}^{(m)} \to \mathcal{Y}$ that maps each data point, or *instance*, x of the input space into an output, or *outcome*, $y = b(x)$. Given a set of instances $X \subseteq \mathcal{X}^{(m)}$, we use $Y = b(X)$ as a shorthand to denote the set of outcomes $Y = \{b(x)|x \in X\}$. A specific instance $x \in \mathcal{X}^{(m)}$ is represented by set of m *features* whose values identifies the data point in the input space.

In the case of a black box model, the function b is either unknown or uninterpretable by humans. Considering interpretability as *the degree to which a human can understand the cause of a decision* [241], a function b can be uninterpretable by humans for several reasons. Consider for example a black box model that takes as input 28×28 pixel dermatology images of moles and outputs whether the mole is benign or malignant. In this case, the input space $\mathcal{X}^{(m)}$ is the set of all possible 28×28 pixel images of moles, an instance $x \in \mathcal{X}^{(m)}$ is one of these images, the m features are the pixels ($m = 28 \times 28$) whose values identify the image, and the output space is $\mathcal{Y} = \{-1, +1\}$ that encodes whether the depicted mole is benign (outcome $y = +1$) or malignant (outcome $y = -1$). The black box model might be uninterpretable because b might be a highly nonlinear function of the pixel values of the image that do not have any semantic meaning for humans. Another possibility is

that the number of parameters of the function b is too high for the limited capacity of human cognition [223].

In the context of the *outcome explanation problem*, the explanation of a black box outcome $y = b(x)$ is generally provided in terms of the input features that generated such outcome. However, the input features are not always interpretable by humans. For example, in the context of dermatology images, when trying to classify a mole as malignant or not, doctors do not reason in terms of single pixels. To generate human-understandable explanations it is therefore important to provide them in terms of a *human-interpretable domain* $\mathcal{E}$ which is a transformation of the original features space and needs to be defined on a case-by-case basis. The explanation is therefore defined as an instance in the human-interpretable domain $e \in \mathcal{E}$.

There are many ways to obtain local explanations. Local XAI techniques can be either *specific* for a particular black box model (e.g., they can only be applied to a certain kind of neural networks) or *agnostic* with respect to it. The model-specific approaches usually involves a process that can only be applied to one particular kind of of ML models. For example, a XAI technique for images might be limited to convolutional neural networks because it needs to perform a backpropagation process to obtain an explanation for the network's outcome [337, 319]. In this thesis we focus on the *model-agnostic* outcome explanation problem. In particular, we focus on exploiting *interpretable* models $c : \mathcal{X}^{(m)} \rightarrow \mathcal{Y}$ (defined in greater detail in chapter 2) that are able to mimic the black box model behaviour in a *neighborhood* $\tilde{X} \subset \mathcal{X}^{(m)}$ of the instance we want to explain, i.e. ideally $c(x) = b(x)\ \forall x \in \tilde{X}$. How this local neighborhood is defined is what differentiate most of the local XAI techniques and is discussed in greater details in section 5.1.2.

The intuition behind many of the *local* approaches to explanation is that even if the decision boundary learned by the black box in the feature space can be arbitrarily complex, locally, it can always be faithfully approximated by a simpler, more interpretable model c, also called *local surrogate model* (a simple representation of

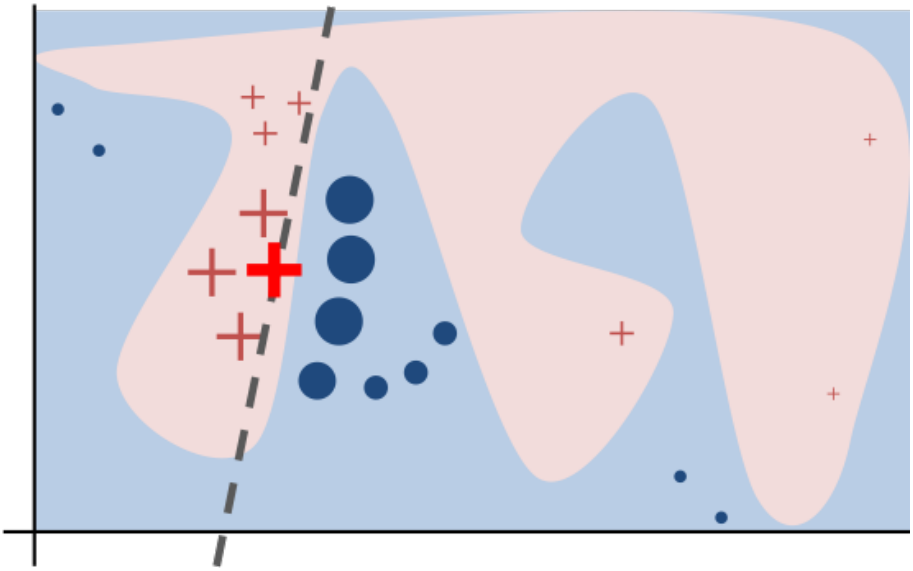


Figure 5-1: Toy example in 2 dimension of a complex decision boundary for a binary classification task (from [302]). The red cross is the instance to be explained, the dotted line is the local surrogate model (linear in this case).

this intuition is shown in Figure 5-1). These local interpretable ML models might be linear models, decision trees or any inherently interpretable model. Since their goal is to mimic the decision-making process of the black box, these local interpretable models are trained using the instances of the neighborhood $\tilde{X}$ of x and the related set of black box outcomes $Y = b(\tilde{X})$. We write $c = f(b, x)$ to indicate that the local predictor is the result of a process $f(\cdot, \cdot)$ that exploits both the instance x whose outcome $y = b(x)$ we want to explain, and the black box model b. In this thesis, the process $f(\cdot, \cdot)$ includes the definition of the neighborhood $\tilde{X}$ of the instance x whose outcome needs an explanation and training the local model c. Finally, the local interpretable model c is used to generate and explanation in the human-understandable domain $e \in \mathcal{E}$. We write $e = \varepsilon(c, x)$ to identify the process $\varepsilon(\cdot, \cdot)$ that extracts an explanation for the instance x from the local classifier c.

5.1.2 The local neighborhood

As explained in the previous section, the local predictor c is usually trained on a feature space *local neighborhood* $\tilde{X}$ of the instance whose outcome we want to explained. One key aspect that differentiate the XAI methods that employ this approach to

local explanations is how they generate such a neighborhood [122, 139, 358, 207]. Indeed the neighborhood should be both *local* and *expressive* enough to allow the interpretable model c to learn the black box model's local behavior. The locality of the neighborhood is usually defined by a distance function $d(x, x')$ that ensures that the instances $x \in \tilde{X}$ are *close* to the instance whose outcome we want to explain. The closer we are to the instance to explain in the feature space, the greater the chances of the local decision boundary being simple enough to be learned by c (as exemplified in figure 5-1). At the same time, such a neighborhood should be *expressive*, i.e., $Y = b(\tilde{X})$ should contain different outcomes. For example, in the case of a binary classification problem where $\mathcal{Y} = \{-1, +1\}$, the instances of the neighborhood should have both outcomes $y = -1$ and $y = +1$. Indeed, if the instance we want to explain is not close to the local black box decision boundary, there is the risk that its neighbors will all have the same outcome. In this case we would have a training set for c that does not contain the local decision boundary of the black box.

In order to generate such an optimal neighborhood, many XAI methods perform a local synthetic augmentation, i.e. they create synthetic local instances close to the one to be explained. Indeed, synthetically increasing the local density of the feature space allows the surrogate model to better mimic the local behaviour of the black box [86]. Ideally, these synthetic instances should be drawn from the true underlying local distribution $\mathcal{X}_l^{(m)}$. Unfortunately, this distribution is generally unknown, therefore the synthetic instances are created extrapolating information on $\mathcal{X}_l^{(m)}$ from a set of available instances $x \in \mathcal{X}^{(m)}$ or from the instance to be explained itself. The problem of generating meaningful synthetic instances has been tackled in many different ways.

In [302] the authors present LIME (Local Interpretable Model-agnostic Explanations), where they fit a local linear model on a synthetic neighborhood generated by perturbing a human-understandable representation $\mathcal{E}$ of the instance to be explained x. The locality of the generated synthetic neighbors is preserved by weighting each instance according to its distance from x. However, the *expressiveness* of the re-

sulting neighborhood is not guaranteed. LIME can be applied to many data types (images, text, tabular) and it employs *occlusion* as its main perturbation strategy, e.g. in the case of images, it creates synthetic instances by randomly replacing the pixel value of some parts of the image with the average pixel value of the overall image. LIME adopts a slightly different approach to perturbation when applied to tabular data. In this latter case the synthetic neighborhood is generated by perturbing x by sampling from a normal distribution having mean and standard deviation of the distribution of that feature in the training set. In [397] the authors present a similar approach to [302], however, they do not create a synthetic neighborhood, instead they first partition the training data using a hierarchical clustering algorithm, then they use a k-nearest neighbors algorithm to select one of these clusters as the set of local instances to fit the surrogate model. Other approaches use a genetic algorithm to create the synthetic neighborhood [186, 143]. In particular, in [143] the authors present an explainability methods which employs a genetic algorithm to generate a neighborhood which is then used to train a decision tree from which a decision and a counterfactual rule are extracted. This approach guarantees the *expressiveness* of the local neighborhood by including a term that optimize for a varied set out local outcome in the fitness function of the genetic algorithm. Other approaches focus on creating a local and expressive neighborhood by perturbing the instances in a latent space [141, 327, 179, 358]. Pros and cons of these approaches in comparison to ours are discussed in section 5.5.

We choose to empirically extrapolate the local distribution of features from a set a set of neighbors of the instance to be explained. Our hypothesis is that perturbing the features according to such empirical distributions should ensure locality. In particular, in section 5.3 and 5.4 we will present two solutions to the *outcome explanation problem* that propose two new approaches to generate the local neighborhood: one tailored to ensure both *locality* and *expressiveness* in the case of multi-label outcomes, and one that exploits feature space semantic knowledge. Both of them first define locality around the instance to be explained according to a *distance function* and use this function to find a set of instances close to it from a set of known in-

stances. Then they perturb such instances to create a synthetic local neighborhood around the instance to be explained which becomes the training set for a decision tree acting as the local interpretable model c. A local explanation is then extracted by the decision tree trained on the synthetic neighborhood. The explanation is in the form of a decision rule $r{=}(p \rightarrow y)$ including in its premise p the split conditions on the path from the root to the leaf node that is satisfied by the instance x. We chose to express the explanation in the form of a decision rule, e.g. a logic-based statement of the type *IF ... THEN*, because using logic allows the user to reason over the explanation [274, 16].

5.1.3 Multi-label classification tasks

Multi-label classification is the task of learning to assign a set of non-mutually exclusive labels to each instance in the feature space. These tasks are quite common in healthcare [356, 379]. For example when there is the need to simultaneously predict the risk of several chronic diseases [218, 409, 127, 117], when trying to classify unknown genes functional expressions [31, 77], when building a clinical algorithm to predict the diagnoses and medications order of patient's future visit [72, 305], when trying to learn multiple indicators of early-stage diseases [75] or when performing clinical text categorization or annotation [32, 96, 403]. More formally, the multi-label classification task can be defined as [405]:

Definition 1 (Multi-label classification task) *The* multi-label classification task *consist in learning a function* $b{:}\mathcal{X}^{(m)}{\rightarrow}\mathcal{Y}^{(l)}$ *which maps data instances* x *from a feature space* $\mathcal{X}^{(m)}$ *with* m *input features to a decision vector* y *in a label space* $\mathcal{Y}^{(l)}{=}\{0,1\}^l$.

An instance x consists of a set of m attribute-value pairs (a_i, v_i), where a_i is a feature (or attribute) and v_i is a value from the domain of a_i. The domain of a feature can be continuous or categorical. Note that, $y_i{=}1$ if the i^{th} label is associated with the instance x, $y_i{=}0$ otherwise.

There are several approaches to solve a multi-label classification task. The most intuitive one is the *binary relevance* approach [46], which consists in decomposing the multi-label classification into multiple independent binary classification tasks. However, this approach is not an easily scalable solution since the output space dimensionality grows exponentially with the number of potential classes (if there are l possible classes, then the output space is 2^l). Furthermore, most of the *binary relevance* approaches do not take into consideration some important information contained into the potential correlations between the different labels [227, 402]. For this reason, a plethora of other works focus on adapting classical learning algorithms to exploit such information [299, 357, 404, 297, 413]. Particularly relevant to the work in this thesis are multi-label decision trees [77, 93, 368], which are used as local surrogate model to extract an explanation.

As mentioned in section 5.1.2, such surrogate model is trained on a *synthetic local neighborhood* of the instance to be explained. The synthetic neighborhood is created to increase the number of instances in the vicinity of the point of interest, then these instances are labeled using the black box, and finally, they are used as a training set for the interpretable classifier. This procedure is done to help the surrogate model better understand the black box's local decision boundary. For this reason, we want to avoid the creation of a synthetic neighborhood that contains instances classified with the same label, i.e., the neighborhood should be *expressive enough* to contain the relevant part of the local decision boundary. A common approach to build such an *expressive* neighborhood is to make sure that all the labels are represented by the selected neighbors. However, while in the case of a binary and multi-class task, creating a local and expressive neighborhood is a straightforward task, in the multi-label case, a trade-off exists. Indeed, since the output space grows exponentially, not all combinations of labels preserve neighborhood locality, i.e., they might not be close to the instance to be explained. We study this trade-off in section 5.3 and in particular in section 5.3.2.

5.1.4 Sequential Data

Sequential data is any data that contains instances whose representation implies some sort of order. Some examples of sequential data are text (an ordered sequence of words or characters), video, DNA (a sequence of nucleotides), the history of consumers' purchases or spatial trajectories. This kind of data is also quite common in healthcare. For example, patients' clinical histories can be represented as sequences of clinical events over time, disease progression can be represented as sequences of symptoms and conditions, medications histories are inherently sequential, and finally, physicians' clinical notes are sequences of words describing the patient encounter. Sequential data has been fed into AI algorithm to perform disparate healthcare-related tasks such as next-visit diagnoses prediction [72, 224], information extraction from clinical notes [175], prediction of hospital readmission [293], and prediction of risk of life-threatening conditions such as heart failure [183], suicidal tendencies [276] and glaucoma [217].

Most of the explainability approaches related to sequential data modeling are model-specific, i.e., they can be applied only to some types black box model. The most popular technique focus on adding an *attention mechanism* [26, 360] to a sequential model and use the attention weights as a form of explanation [74, 231, 393, 27], however recent works have highlighted how this kind of explanation might lack consistency [176, 323, 50] and that attention should not be used as an explanation. Other approaches related to sequential data modeling focus on understanding the internal behavior of the black-box under study [387, 191].

However, none to few of the explainability methods for sequential data present in the literature are agnostic with respect to the black box [330]. This is probably due to the fact that, for most sequences, is not straightforward to pre-define a sequential human-interpretable domain $\mathcal{E}$ that solves the outcome explanation problem (detailed in section 5.1.1). Furthermore, there are no off-the-shelf inherently interpretable sequential models able to act as local surrogate models. We studied this problem in section 5.4 and in particular in section 5.4.3. A similar work (which

has been published after ours) solves this problem by proposing *TimeSHAP* [37], a variation of the popular SHAP method [228] that, similarly to our approach, performs sequence perturbations. However, our approach takes also into consideration the semantic 5.1.5 relationships among the perturbed features through the use of ontologies, as better explained in section 5.4.3.

5.1.5 Ontology-linked data

There is no agreed upon formal definition of what an ontology is. The word *ontology* might mean different things for different communities and in different contexts. In Computer Science, an ontology is usually defined as *an explicit specification of a conceptualization* [137], or more specifically:

> "An ontology is a formal, explicit specification of a shared conceptualisation. A *conceptualisation* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. *Explicit* means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. *Formal* refers to the fact that the ontology should be machine readable, which excludes natural language. *Shared* reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group". [341]

In other words, an ontology is a structured, machine-readable representation of the knowledge pertaining a specific aspect of a domain. In order to be machine-readable, the language used to encode the knowledge must have formal properties that are well understood, which means that usually ontologies are specified using logic-based languages that can also be used to perform computational inference through automated reasoning.

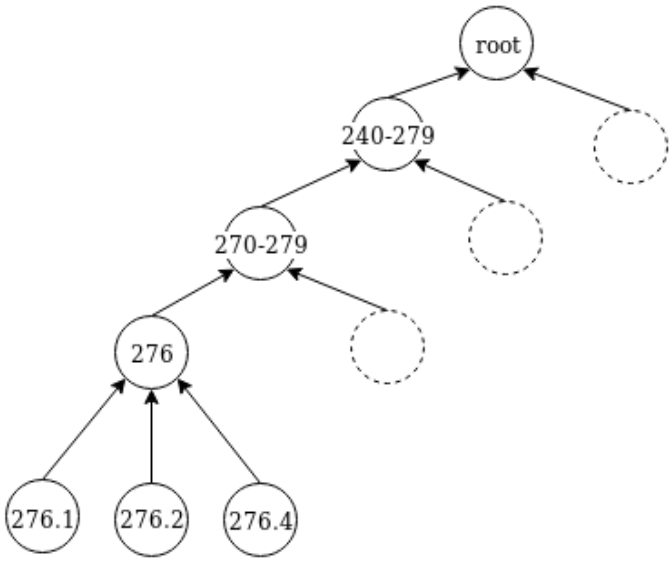
In the literature, two kinds of ontologies are distinguished; *lightweight* and *heavyweight* ontologies [83]. Lightweight ontologies [132] define a set of vocabulary terms of the domain of interest, also referred to as *concepts*, and encode all their relevant properties and relationships. Heavyweight ontologies also provide constraints on the use of the concepts and their relationships and therefore are able to model the knowledge in a deeper way. Ontologies can be visualized using graphs, where the nodes are the concepts and the links are the relationships among the concepts.

When data are *ontology-linked* its items can be linked to the concepts represented in an ontology, e.g., the words in a text documents can be mapped to WordNet [240], a lexical ontology containing relatioships between words in multiple languages. The presence of ontology-liked data is widespread in the medical and biological fields. A medical ontology might capture different aspects of the field of medicine. For example, it might represent the knowledge of anatomy and physiology or it could encode medical terminology. Some notables examples are the Disease Ontology [315], the Open Biomedical Ontology (OBO) [334], the Diabetes mellitus Diagnosis Ontology (DDO) [104], the Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) [97] and the Unified Medical Language System (UMLS) [42]. In the following we focus on the ICD-9 codes ontology since it is the one that we mostly used in this thesis.

ICD-9 codes ontology

The *International Classification of Diseases* (ICD) is the standard for the reporting and coding of diseases and health conditions [383]. In its Ninth Revision, Clinical Modification (ICD-9-CM) the codes have an alphabetic or numeric first digit, the remaining digits are numeric. Their length can vary from a minimum of three digits to a maximum of five digits. Their structure is the following [58]:

XXX	.XX
Category (digits 1–3)	Etiology (digits 4–5)
	Anatomic site
	Manifestations



ICD-9	description
root	root
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
270-279	Other Metabolic Disorders And Immunity Disorders
276	Disorders of fluid electrolyte and acid-base balance
276.1	Hyposmolality and/or hyponatremia
276.2	Acidosis
276.4	Mixed acid-base balance disorder

Figure 5-2: A representation of a branch of the tree-shaped ICD-9 hierarchical ontology: the root is a general condition *Disease* while its children and grandchildren are increasingly more specific conditions.

So the first 3 digits identify the category of the diagnosis (e.g. *infectious and parasitic diseases*, *endocrine, nutritional and metabolic diseases, and immunity disorders*) while the last 2 digits identify the etiology or the anatomic site of the diagnosis.

The set of hierarchical relationships between these codes constitues a *taxonomy*, i.e. a *lightweight* ontology containing concepts related by the simple relationship type *"is-a"*, e.g. *250: Diabetes mellitus* "is-a" *249-259: Diseases Of Other Endocrine Glands* "is-a" *240-279: Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders* "is-a" *ROOT: Disease*. A branch of the ICD-9 ontology is represented in Figure 5-2.

ICD codes' main use is to share health information in a structured way. In particular, they are used to share patients' clinical history across hospitals, to monitor

diseases' prevalence and incidence, to evaluate hospital performances, and to fill the claims for health insurance reimbursement. ICD codes also allow for data-driven health policies, and recently they have been exploited to build clinical decision-support-systems (DSS) based on Machine Learning (ML) models [72, 63, 160]. Most of ML-based DSS trained on ICD codes assume that these are a good proxy for the patient's actual health status. However, ICD codes can misrepresent such status because of many potential sources of error in the translation of the patient's actual disease into the respective code (see [261] for a complete description of these sources of error).

ICD-9 codes can also be mapped in a smaller set of codes using the *Clinical Classifications Software* (CCS) [106] which is a diagnosis and procedure categorization scheme that collapses them into a smaller number of clinically meaningful categories.

Ontological similarity measures

Having ontology-linked data allows to calculate a set of *ontological similarity measures* [7] among the data items, i.e., finding data points that are semantically similar. This is relevant for the creation of a neighborhood to solve the outcome explanation problem (detailed in section 5.1.1). Such measures of similarity are built starting from measures of similarity among the concepts of the ontology under study. These similarity measures use the relationships encoded in the ontology to determine if two concept are *semantically* similar, which intuitively means that they have similar meanings in the context provided by the ontology.

Two main approaches to semantic similarity are present in the literature [284]: those based on *Information Content* (IC) and those based on path measures on the ontology graph representation. IC-based measures of similarity adopt a probabilistic approach based on the frequency of occurrence of concepts in the data and are based on the intuition that similar concepts have similar degree of informativeness [300], while path-based measures are based on distances that takes into consideration the edges that connect two nodes in the graph representation of the ontology, e.g., edge

counting [291] and weighted edge counting [199, 211]. Path-based measures are particularly well suited for taxonomies (hierarchical ontologies with only "is-a" relationships) because for most of them concepts that are higher in the hierarchy, i.e. they are close to the root, represents more general concepts than those close to the leaves. Anyway, it has been shown that these two approaches to ontological similarity measures are related and some of them can be derived from the same generalized model of similarity [89].

We exploit one path-based ontological similarity measure, the Wu and Palmer similarity [389], to study the creation of an ontological neighborhood to extract local explanations that solve the outcome explanation problem with an agnostic approach in section 5.4 and in particular in section 5.4.3.

5.2 Main contribution

This chapter is a based on two of our papers:

- Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pages 97–110. Springer, 2019
- Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 629–639, 2020

In the first paper we presented **MARLENA** (**M**ulti-l**A**bel **R**ul**e**-based **E**xpla**NA**tions), a model-agnostic XAI methodology to address the outcome explanation problem in the context of multi-label black box outcomes. MARLENA explains an individual black box decision in three steps. First, it generates a synthetic neighborhood around the instance to be explained using a strategy suitable for multi-label decisions. It then learns a decision tree on such neighborhood and finally derives from it a decision rule that explains the black box decision. Our experiments show that MARLENA performs well in terms of mimicking the black box behavior while gaining at the same time a notable amount of interpretability through compact decision rules, i.e., rules with limited length.

Building on the insights we gained from the experiments carried out in this first paper, we developed **Doctor XAI**, a model-agnostic technique that is suitable for multi-label black box outcomes and it is also able to deal with sequential inputs. More importantly, Doctor XAI exploits the medical domain knowledge encoded in ontologies in its explanation process. In particular, Doctor XAI performs a querying process of the model which is semantically meaningful for a domain expert. We show that exploiting the temporal dimension in the data and the domain knowledge encoded in the medical ontology improves the quality of the mined explanations.

5.3 MARLENA: multi-label black box outcome explanation

In this section, we present **MARLENA** (**M**ulti-l**a**bel **R**ul**e**-based **E**xpla**NA**tions) as an agnostic solution to the multi-label black box outcome explanation problem. Given any kind of multi-label black box classifier b and a specific instance x labeled with outcome y by b, MARLENA first generates a set of synthetic neighbors close to x using an ad-hoc strategy, then uses b to label such neighbors and uses them to train a multi-label decision tree classifier. Finally, it explains $y = b(x)$ by extracting a decision rule containing all the split conditions of the decision tree that lay on the path from the root to the leaf that matches x. For the generation of the neighborhood of x, we test two alternative strategies based on the idea of generating neighbors close to x both in the features and in the label space. We validate MARLENA with experiments on real datasets to quantitatively assess its accuracy in mimicking b and the complexity of its explanations.

5.3.1 Problem definition

As detailed in section 5.1.1, given a black box classifier b, human interpretable domain $\mathcal{E}$, and an instance x, the *outcome explanation problem* consists in providing an explanation $e \in \mathcal{E}$ for the decision $y = b(x)$.

We address this problem in the specific case in which the black box is a *multi-label classifier* $b{:}\mathcal{X}^{(m)} \rightarrow \mathcal{Y}^{(l)}$. Our approach is based on the idea of learning an interpretable classifier c that reproduces and accurately emulates the *local* behavior of the black box. An explanation for the decision is then derived from c.

By *local*, we mean that we focus on the behavior of the black box in the *neighborhood* of the specific instance x, without aiming at providing a description of the overall decision-making process of the black box. The neighborhood of x has to be generated as part of the explanation process.

We assume that some knowledge is available about the feature space $\mathcal{X}^{(m)}$, like the ranges of admissible values for the domains of the features and, like in this work, the (empirical) distribution of the features. Nothing is instead assumed about the process of constructing the black box b. Let us define the problem of outcome explanation through interpretable models:

Definition 2 (Explanation through interpretable models) *Let $c = f(b, x)$ be an interpretable classifier derived from the black box b and the instance x using some process $f(\cdot, \cdot)$. An explanation $e \in \mathcal{E}$ is obtained through c, if $e = \varepsilon(c, x)$ for some explanation extraction process $\varepsilon(\cdot, \cdot)$ which involves c and x.*

In the next section we will describe the process $f(\cdot, \cdot)$ we propose for obtaining an interpretable classifier c. Similarly to [143], we adopt as explanation a *decision rule* (simply, a rule) r of the form $p \rightarrow y$ describing the reason for the decision value $y = c(x)$. The decision y is the *consequence* of the rule, while the *premise* p is a boolean condition on feature values.

Definition 3 (Local explanation) *Let x be an instance, and $c(x) = y$ be the decision of an interpretable multi-label classifier c. A local explanation e is a a decision rule $r = (p \rightarrow y)$ consistent with c and satisfied by x.*

We assume that p is the conjunction of split conditions sc of the form $a \in [v_1, v_2]$, where a is a feature and v_1, v_2 are values in the domain of a extended with $\pm\infty$.

An instance x *satisfies* r, or r *covers* x, if the boolean condition p evaluates to true for x, i.e., if $sc(x)$ is true for every $sc \in p$. For example, the rule

$$\begin{aligned} r = \{ & 60 < age \leq 70, \\ & BMI > 36.2, \\ & hyperglycemia = Yes, \\ & \rightarrow [Diabetes, Hypertension, Hypothyroidism] \end{aligned} \tag{5.1}$$

is satisfied by:

$$x_0=\{age=63, BMI=36.5, hyperglycemia=Yes\}$$

while is not satisfied by:

$$x_1=\{age=65, BMI=35, hyperglycemia=No\}$$

We say that r *is consistent* with c, if $c(x)=y$ for every instance x that satisfies r. Consistency means that the rule specifies some conditions for which the classifier makes a specific decision. When the instance x for which we have to explain the decision satisfies p, the rule $p \rightarrow y$ represents *a motivation for taking* a decision value, i.e., p locally explains why b returned y. Therefore, a solution to the problem will consists of:

1. Computing an interpretable predictor c for a black box b and an instance x, i.e., designing function $f(\cdot,\cdot)$ according to Definition 2;
2. Deriving a local explanation e from c and x, i.e., defining the explanation extraction process $\varepsilon(\cdot,\cdot)$ according to Definition 3.

Let us consider as an example the following explanation for the diagnoses prediction of a patient:

$$\begin{aligned} e =\{&60 < age \leq 70,\\ &BMI>36.2,\\ &hyperglycemia=Yes,\\ &insulin=Up,\\ &systolicpressure=150/100mmHg\}\\ &\rightarrow[Diabetes, Hypertension, Hypothyroidism] \end{aligned} \tag{5.2}$$

The meaning of this explanation is that the diagnoses of *diabetes*, *hypertension* and *hypothyroidism* are predicted by the black box because the patient is obese (BMI>36.2), his systolic pressure is high, his age is in the $[60, 70)$ range and his

blood test results show high levels of sugar (hyperglycemia) and insulin. For the sake of clarity, we only show the diseases that have been predicted by the black box, which correspond to non-zero elements of the binary label vector $y \in \mathcal{Y}^{(l)} = \{0,1\}^l$.

5.3.2 Neighborhood Generation

The goal of this phase is to create a local training set Z for the local surrogate model c which contains the local decision behavior of the multi-label black box b. Locality means that the instances of such a training set should be neighbors of the instance to explain x. In addition to being *local*, this training set should also be *expressive* enough to include instances with both decisions equal to $b(x)$, i.e. $b(z)$=$b(x)$ and decisions different from $b(x)$, i.e, $b(z)\neq b(x)$. Given the multi-label setting, which grows exponentially the number of possible combinations of labels satisfying $b(z)\neq b(x)$, we can not include in Z all the possible ways in which the black box decision differ from $y = b(x)$. Our hypothesis is that we can build a local neighborhood Z that is both *expressive* and *local* by considering locality both in the features space $\mathcal{X}^{(m)}$ and in the label space $\mathcal{Y}^{(l)}$.

In order to do so we propose two approaches that first identify a *core real neighborhood* X^* of x using a distance function that takes into consideration both feature space and label space. This set of instances is selected from a set of *known instances* $\hat{X}\in\mathcal{X}^{(m)}$ that may be a set of instances of the training set, a set of instances to be explained or in general, a set of instances belonging to the same domain of x (in our experiments, we setup $\hat{X}$ as the instances to explain in the test set). Once identified this core of real instances, we use them to derive the empirical local distributions of features around x, and then we randomly generate the set of *synthetic neighbors* Z according to these distributions.

We tested two ways of considering locality both in the features space and in the label space to which correspond two different distance functions. In the following, we describe **MARLENA-m** and **MARLENA-u**, the two resulting versions of MARLENA.

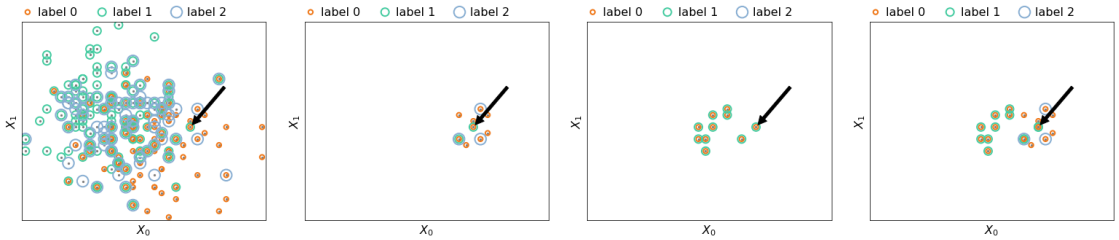


Figure 5-3: *(1st plot)* 2D feature space with multi-label instances having 3 types of different labels, the arrow points out the instance to explain. *(2nd plot)* MARLENA-m selects αk neighbors in the feature space. *(3rd plot)* MARLENA-m selects $(1-\alpha)k$ neighbors in the latent space. *(4th plot)* The resulting *mixed neighborhood* obtained merging merge the previous sets of instances.

MARLENA-m: mixed neighborhood

This method introduces a parameter $\alpha \in [0, 1]$ which allow us to set the percentage of neighbors that we want to take from the features space. MARLENA-m selects from the given instances $\hat{X}$ a core of k real neighbors $X^* = X_f \cup X_l$, where $k = k_f + k_l, \quad k_f{=}\alpha k$ and $k_l{=}(1{-}\alpha)k$

- The set X_f is composed of the k_f instances $\hat{x} \in \hat{X}$ closest to x with respect to the feature space $\mathcal{X}^{(m)}$, according to a distance function $d_f(x, \hat{x})$
- The set X_l comprises the k_l instances $\hat{x} \in \hat{X}$ closest to x with respect to the target space $\mathcal{Y}^{(l)}$, i.e., the black box decision, according to a distance function $d_l(b(x), b(\hat{x}))$.

The parameter α is then an hyperparameter of the MARLENA-m approach to be set according to the task at hand. Our hypothesis is that such parameters helps exploring the expressiveness of the created neighborhood (instances in X_l which are close to x with respect to the black-box decision are not necessarily close to x in the feature space). Low values of α could bring to the generation of a sparse real core neighborhood in the feature space. Figure 5-3 shows a graphical representation of mixed neighborhood generation starting from a sample dataset with three different labels (left most plot).

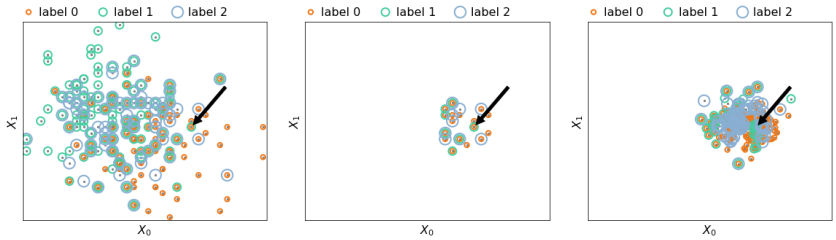


Figure 5-4: *(1st plot)* 2D feature space with multi-label instances having 3 types of different labels, the arrow points out the instance to explain. *(2nd plot)* MARLENA-u selects k neighbors using the distance function which combines distances in the features and in the label space generating in one single step the *unified neighborhood*. *(3rd plot)* MARLENA-u perturb the core of k neighbors generating a dense *synthetic neighborhood* around the instance to explain.

MARLENA-u: unified neighborhood.

This method selects from the given instances $\hat{X}$ a core of k real neighbors X^* as the k instances $\hat{x} \in \hat{X}$ closest to x with respect to both the feature space $\mathcal{X}^{(m)}$and the target space $\mathcal{Y}^{(l)}$, according to a distance function $d_u(x, \hat{x}, b)$ which combines d_f and d_l:

$$d_u(x, \hat{x}, b) = \frac{m}{m+l} \cdot d_f(x, \hat{x}) + \frac{l}{m+l} \cdot d_l(b(x), b(\hat{x}))$$

.

For an example see Figure 5-4 *(1st plot)* and *(2nd plot)*.

Both approaches are parametric with respect to the distance functions $d_f(\cdot, \cdot)$ and $d_l(\cdot, \cdot)$. Since we have binary vectors with length l, in the target space we use the Hamming distance as $d_l(\cdot, \cdot)$. On the other hand, in the feature space we account for the presence of mixed types of features by a weighted sum of the Hamming distance [346] for categorical features, and of the normalized Euclidean distance[1] for continuous features. Thus, assuming s categorical features and $m - s$ continuous ones, we use:

$$d_f(x, \hat{x}) = \frac{s}{m} \cdot Hamming(x, \hat{x}) + \frac{m-s}{m} \cdot nEuclidean(x, \hat{x})$$

Beside α for MARLENA-m, both approaches have two other hyperparameters:

[1] http://reference.wolfram.com/language/ref/NormalizedSquaredEuclideanDistance.html

the number of core real neighbors k and the number of synthetic neighbors Z to generate k_{syn}. All of these parameters need to be set according to the task at hand and to the characteristics of the dataset.

Synthetic neighborhood creation

Once identified the set of *core real neighbors* X^* from the available set of instances $\hat{X}$, MARLENA generates a *dense synthetic neighborhood* Z of x using the empirical features distributions of X^*. In particular, in our experiments, for each continuous features we sampled from a Gaussian $\mathcal{N}(\mu; \sigma^2)$ distribution having mean and variance equal to the ones calculated from the empirical distribution for that feature in X^*. Similarly, for each categorical feature we randomly sampled one value according to its empirical frequency in X^*. However, we stress that our approach can be employed with other kinds of sampling. We can see the an example of the resulting synthetic neighborhood Z resulting from a *unified core real neighborhood* in Figure 5-4 *(3rd plot)*.

Rule-based explanations

Given the synthetic neighborhood Z of x, the second step is to build an interpretable classifier c trained on the instances $z \in Z$ labeled with the black box decision $b(z)$.

Such a classifier is intended to mimic the behavior of b locally in the Z neighborhood. **MARLENA** adopts multi-label decision tree as interpretable classifier c as it makes easy the explanation extraction. Indeed, given the multi-label decision tree c, we derive the decision rule representing the explanation as a root-leaf path in the tree, i.e., the decision rule $r = (p \rightarrow y)$ is formed by including in p the split conditions on the path from the root to the leaf node that is satisfied by the instance x, and setting $y = c(x)$. By construction, the rule r is consistent with c and satisfied by x.

5.3.3 Experiments

In this section, we describe the experiments we carried out to evaluate the performance of **MARLENA**. We first present the experimental setup and then we show the results of our analyses which prove that the proposed multi-label local approach is more effective than a global one. We study the effect of the neighborhood generation parameter α on **MARLENA-m** performance, and we provide a qualitative and quantitative evaluation of the multi-label explanations. **MARLENA** was developed in Python[2], we used the `sklearn` implementation of the multi-label decision tree as interpretable classifier.

Datasets

We ran experiments on three real-world multi-label tabular datasets: *yeast* [105], *woman*[3] and *medical* [278]:

- The *yeast* dataset is a collection of yeast microarray expressions and phylogenetic profiles which can be used to learn the yeast gene functional categories. One row of this dataset represents a gene, and the labels are its associated functional classes. Each gene might belong to more than one functional class.
- The *woman* dataset contains survey data about women health-care requirements gathered by a US non-profit organization. One row of this dataset contains the questionnaire replies of one woman concerning her demographics, pregnancies, family planning, use of health care services, and medical insurance. The labels of this dataset are the health-care requirements.
- The *medical* dataset contains a corpus of fully anonymized clinical text. Each document in the corpus is associated with a set of ICD-9 codes which represents the diagnosis associated with the clinical report. To each report might be assigned several ICD-9 codes.

[2]Source code, datasets, and the scripts for reproducing experiments are publicly available at https://github.com/riccotti/ExplainMultilabelClassifiers

[3]https://www.kaggle.com/ravikrishnareddy/multi-label-classification

Dataset	instances	features	labels	avg. labels	RF	SVM	MLP
yeast	2,417	117	14	4.24	.62	.62	.64
women	14,644	44	14	3.53	.71	.72	.71
medical	978	1449	45	1.25	.37	.79	.77

Table 5.1: Real health-related dataset information and black box performance (F1-measure).

The *woman* dataset includes both categorical and continuous features, the *yeast* only continuous features and the *medical* dataset contains only binary features that represent the presence or absence of each word in each document. Details of the datasets after missing values correction [4] and black box performance are reported in Table 5.1. To train the black boxes, we randomly split the *yeast* and *woman* dataset into a training and a test set containing respectively 70% and 30% of the instances. For the *medical* dataset we use the partitioning described in the related paper [278].

Black box classifiers

After the training phase we used the black boxes to classify the instances in the test set, denoted by X, and we used the **MARLENA** approach to explain such decisions. We denote by $\hat{Y}$ the decisions provided by the black box b on X, and with Y the decisions provided by the explainer c. We underline that the black box performance is not the focus of our work: once the black box is trained on the training set we forget about the real label associated with each instance and we use the black box labels as target labels for the training of the decision tree.

We experimented the following predictors as black boxes: Random Forests (**RF**), Support Vector Machines (**SVM**), and Multi-Layer Perceptron (**MLP**)[5]. For each black box, we performed an hyper-parameters tuning using a five-fold cross-validation and a randomized search over a grid of parameters on the training set.

[4] We replace the missing values with the mean for continuous variables and with the mode for categorical ones. We remove the features with more than 40% of missing values.

[5] Implementations are those of `scikit-learn` library.

Evaluation Measures.

We adopt the following metrics to evaluate **MARLENA**'s performance in explaining black box decisions and in mimicking its local behavior. Aggregated values of *fidelity* and *hit* are reported by averaging them over the set X.

- ***fidelity***$(Y, \hat{Y}) \in [0, 1]$. It compares the decisions of the interpretable classifier c to those of the black box b on the set X. The *s-fidelity* measures the performance on the synthetic neighborhood, $X{=}Z$. The *r-fidelity* measures the performance on the core real neighborhood, $X{=}\hat{X}$. It answers the question: "how good is c at mimicking b in a neighborhood of x?". We measure it using the F1-measure [346].

- ***hit***$(y, \hat{y}) \in [0, 1]$. It compares the prediction of c and b on the instance x under analysis. We use the simple match similarity to evaluate it, i.e., $1 - hamming(y, \hat{y})$. $hit(y, \hat{y}) = 1$ means that c correctly identifies all the labels returned by b, a value between 0 and 1 means that some labels are misclassifed.

5.3.4 Results

In this section we discuss the results of the experiments set up in previous section. For both neighborhood generation approaches *mixed* and *union*, we set the size of the synthetic neighborhood as $k_{syn} = 1000$, and the size of the *core real neighborhood* X^* is computed from the size of the set of known instances as follows: $k = \frac{1}{2}\sqrt{\hat{X}}$

The impact of α on MARLENA-m

We perform several experiments to assess how **MARLENA-m** performance are impacted by the neighborhood generation parameter α. We measure *r-fidelity* and *hit* for different values of α, the results are show in figure 5-5. We observe that, contrary to our hypothesis, the value of α does not have a noticeable impact on the **MARLENA-m** performance. Therefore, we can safely set $\alpha{=}0.7$ for the following analyses, this guarantees the locality in the feature space of the core of real instances selected to generate the synthetic neighborhood. We recall that high values of α favorite neighbors close to x in the feature space.

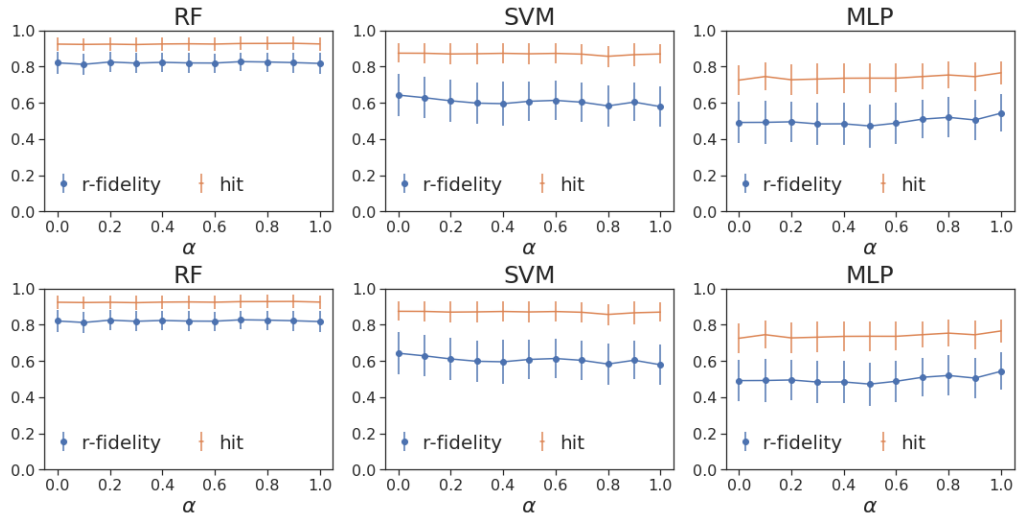


Figure 5-5: Hit and *r-fidelity* varying α for *yeast* and *woman*, upper and lower figure respectively.

Comparison between the two approaches

To understand if one of the two approaches of neighborhood generation performs significantly better than the other, we compare them in terms of their *s-fidelity* and *r-fidelity* on the *woman* and *yeast* datasets. The results are reported in Tables 5.2. We observe that the two approaches have comparable performance, but the *mixed* approach performs slightly better on the synthetic neighborhood.

	s-fidelity		*r-fidelity*	
Black Box	*mixed*	*unified*	*mixed*	*unified*
RF	.94 ± .02	.90 ± .05	.89 ± .09	.87 ± .11
SVM	.91 ± .05	.87 ± .07	.65 ± .20	.68 ± .21
MLP	.93 ± .07	.91 ± .11	.68 ± .22	.68 ± .21

Table 5.2: Fidelity (mean ± stddev) of *MARLENA-m* and *MARLENA-u* on all datasets.

Dataset	yeast		woman		medical	
Black Box	*mixed*	*union*	*mixed*	*union*	*mixed*	*union*
RF	.93 ± .03	.92 ± .04	.94 ± .02	.90 ± .05	.93 ± .06	.90 ± .12
SVM	.84 ± .07	.84 ± .08	.92 ± .03	.88 ± .05	.95 ± .05	.86 ± .14
MLP	.90 ± .05	.90 ± .06	.95 ± .02	.94 ± .04	.80 ± .12	.72 ± .20

Table 5.3: *s-fidelity* (mean ± stddev) of *MARLENA mixed* and *union* for each dataset.

Dataset	yeast		woman		medical	
Black Box	*mixed*	*union*	*mixed*	*union*	*mixed*	*union*
RF	.89 ± .06	.90 ± .06	.89 ± .09	.87 ± .12	.94 ± .09	.97 ± .06
SVM	.86 ± .08	.86 ± .08	.57 ± .16	.60 ± .18	.92 ± .12	.97 ± .06
MLP	.89 ± .06	.89 ± .07	.62 ± .21	.61 ± .19	.81 ± .20	.89 ± .14

Table 5.4: *r-fidelity* (mean ± stddev) of *MARLENA* *mixed* and *union* for each dataset.

We want to highlight that all the reported aggregated performance considers only instances for which an explanation is returned. Indeed, for some instances of the *medical* dataset using the *RF* black box an explanation is not returned. This is due to the fact that, as reported in table 5.1, the performance of the RF on the medical dataset is very poor. This means that when labeling the synthetic neighbors generated by MARLENA, such black box always returns the same label, creating a training set for the surrogate model which does not contain any decision different then the one of the instance to be explained, i.e, the synthetic neighborhood is not expressive enough. When this happens, MARLENA learn a *dummy* decision tree with no internal nodes which always classify instances with the same level. This translates into the creation of a rule with no premises. The creation of a non-expressive local neighborhood might also be due to the fact that some instances are far from the decision boundary.

The impact of the *cohesion* of the dataset

We can also see how the aggregated performance on all datasets show lower values of *r-fidelity* when our methods are used to explain *SVM* and *MLP* decisions. Looking at *r-fidelity* values in Table 5.2, we observe that this behaviour is due to weak performance on the *woman* dataset. This gap of performance among the different datasets is due to the different levels of cohesion of the data points selected in the *core real neighborhood* in the feature space.

In order to quantitatively measure the level of cohesion of each neighborhood, we compute the SSE (Sum of Squared Errors [346]) employing distance function d_f

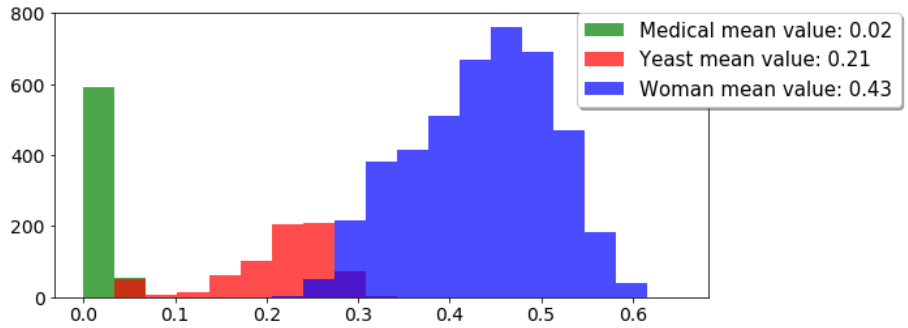


Figure 5-6: Distributions of mean mixed distance among core real neighborhood points.

defined in section 5.3.2. In Figure 5-6 we report the distribution of SSE values, i.e., the mean values of distances among the data points in the core real neighborhoods for each dataset. We observe how the data points in the *woman* dataset are more distant from the center of their neighborhood, compared to those of the other two datasets. This impacts the performance of the methods because selecting data points scattered in the feature space for the core real neighborhood generates a synthetic neighborhood which does not preserve the locality around the instance to be explained.

Quantitative comparison with a global decision tree

Since most of the XAI methods are not directly applicable to the multi-label case, we compared MARLENA against a global approach based on a multi-label decision tree directly trained on all the set of instances that needed explanations labeled by the black box. In particular, we compared the performance of MARLENA to the one of a Global Decision Tree (GDT) for *hit*-performance and rule lenght. We used rule lenght as a measure of comprehensibility of the provided explanation. Indeed, a long rule, i.e., a long explanation, put a high cognitive load on the end user. Therefore shorter rules are preferable.

The results for both the *mixed* and *unified* approaches are shown in Table 5.5 and Table 5.6, respectively. Usually, a global approach to explainability based on approximating the whole black box decision boundary with a single decision tree

Dataset	yeast		woman		medical	
Black Box	MARLENA-m	GDT	MARLENA-m	GDT	MARLENA-m	GDT
RF	.97 ± .05	.98 ± .04	.95 ± .06	.99 ± .04	1.00 ± .01	1.00 ± .01
SVM	.95 ± .06	.93 ± .07	.87 ± .09	.99 ± .03	1.00 ± .01	.99 ± .01
MLP	.97 ± .05	.94 ± .07	.82 ± .13	.99 ± .03	.99 ± .01	.99 ± .01

Table 5.5: Hit performance comparison (mean and standard deviation).

Dataset	yeast		woman		medical	
Black Box	MARLENA-u	GDT	MARLENA-u	GDT	MARLENA-u	GDT
RF	.97 ± .05	.98 ± .04	.94 ± .07	.99 ± .04	1.00 ± .00	1.00 ± .01
SVM	.95 ± .06	.93 ± .07	.87 ± .09	.99 ± .03	1.00 ± .01	.99 ± .01
MLP	.96 ± .05	.94 ± .07	.81 ± .12	.99 ± .03	1.00 ± .01	.99 ± .01

Table 5.6: Hit performance comparison (mean and standard deviation).

should be a weak alternative to a local approach. Indeed, a single decision tree should not have the expressive power to capture the whole complex decision-making process of the black box. However, in this case the the hit performance of the global decision tree (GDT) are high, all above 0.93. This makes the GDT a non-trivial baseline. The high mimicking performance of the global approach is probably due to the characteristics of the selected datasets (small number of instances and generally low degree of multi-labeledness, i.e., the average number of labels per instance). We show that our approaches outperform the global one in mimicking the SVM and the MLP black box on the *yeast* dataset. However, although **MARLENA** in some cases performs worse in terms of hit, it always greatly outperforms the GDT in terms of rule interpretability. Indeed, as shown in Tables 5.7 and 5.8, **MARLENA** always produces explanations (decision rules) with considerable lower number of conditions in the rule premise. The reduction of rule length is really important especially on *woman* dataset.

Dataset	yeast		woman		medical	
Black Box	MARLENA-m	GDT	MARLENA-m	GDT	MARLENA-m	GDT
RF	2.92 ± 2.27	9.09 ± 3.35	4.30 ± .98	13.20 ± 4.56	1.41 ± 1.90	7.70 ± 3.12
SVM	3.29 ± 2.24	5.68 ± 1.47	4.31 ± 1.51	16.30 ± 6.61	5.35 ± 1.67	11.76 ± 4.82
MLP	2.44 ± 1.99	6.70 ± 2.36	2.93 ± 1.17	14.85 ± 6.17	4.58 ± 1.40	10.77 ± 5.40

Table 5.7: Mean rule length and standard deviation comparison between MARLENA-m and GDT.

Dataset	yeast		woman		medical	
Black Box	MARLENA-u	GDT	MARLENA-u	GDT	MARLENA-u	GDT
RF	2.91 ± 2.44	9.09 ± 3.35	4.36 ± 1.19	13.20 ± 4.56	1.80 ± 2.01	7.70 ± 3.12
SVM	3.18 ± 1.99	5.68 ± 1.47	4.36 ± 1.62	16.30 ± 6.61	4.31 ± 2.32	11.76 ± 4.82
MLP	2.70 ± 2.30	6.70 ± 2.36	2.77 ± 1.42	14.85 ± 6.17	4.50 ± 1.75	10.77 ± 5.40

Table 5.8: Mean rule length and standard deviation comparison between MARLENA-u and GDT.

Qualitative comparison with a global decision tree

We now make a qualitative comparison of the explanations provided by **MARLENA-m** and the GDT. We consider explanations for black box behavior on the *medical* dataset since its features are easily comprehensible also by non-experts. What follows is an example of an explanation for the *SVM* black box where both **MARLENA-m** (e_M) and the GDT (e_G) predict the same labels as the black box. In the *medical* dataset the classification task is to map words coming from clinical notes to one or more diagnosis. The following explanations highlights which are the words that influenced more the black box decision with their presence or absence. We highlight words common to both explanations as they probably are the most important for the decision.

$$
\begin{aligned}
e_M = \{&\textit{duplication}{=}0, \textbf{\textit{reflux}}{=}0, \textbf{\textit{hydronephrosis}}{=}1, \textit{normal}{=}1, \textbf{\textit{pyelectasis}}{=}1, \textit{mild}{=}1\} \\
&\rightarrow [\textit{Urinaryincontinence}, \textit{Hydronephrosis}] \\
e_G = \{&\textit{cough}{=}0, \textbf{\textit{reflux}}{=}0, \textit{tract}{=}0, \textit{neurogenic}{=}0, \textbf{\textit{hydronephrosis}}{=}1, \textit{hydroureter}{=}0, \\
&\textit{evaluate}{=}0, \textbf{\textit{pyelectasis}}{=}1, \textit{follow}{=}1\} \\
&\rightarrow [\textit{Urinaryincontinence}, \textit{Hydronephrosis}]
\end{aligned}
$$

We observe that the GDT's explanation is longer and more confusing since it contains words falling outside the context of kidney problems, like *cough*, and generic words such as *evaluate* and *follow*.

5.3.5 Lessons learned

We have proposed **MARLENA** a model agnostic approach to address the multi-label black box outcome explanation problem. Our approach learns a *local* classifier on a synthetic neighborhood generated by a strategy suitable for multi-label decisions. Then, it derives from the interpretable local prediction a meaningful explanation represented by a decision rule, explaining the reasons for the decision. We

have proposed two strategies for the synthetic neighborhood generation that take into consideration the particular structure of the multi-label decision. Our experimentation shows that **MARLENA** presents an acceptable performance in terms of accuracy in mimicking the black box and is able to produce explanations represented by compact rules. In the following, a summary of some key aspects emerging from the analysis:

- Local multi-label decision trees can provide high-quality explanations in terms of fidelity to the black box, hit and compact rules.
- Contrary to our initial hypothesis, considering locality both in the features and in the label space does not seem to make a difference in terms of fidelity and hit. This suggests that it is sufficient to consider neighbors in the feature space.
- Training the surrogate model on a synthetic neighborhood created starting from a set or real neighbors allows to capture the local features distributions.
- Some data points are *harder* to explain because they are far away from the decision boundary. This issue is easily solvable by increasing the number k of core real neighbors until the resulting synthetic ones are labeled with different kinds of black box decisions.
- The results show that different levels of *local data density* highly impact the quality of neighborhood generation. This issue is a drawback of building the training set of the surrogate model starting from the first real neighbors in the dataset: if the dataset is sparse, then the synthetic neighborhood will not preserve locality. This issue is highly dependant on the dataset and can be solved by reducing the initial number of core real neighbors.
- Perturbing independently each feature when creating the local synthetic neighborhood do not consider important relationships between features (thus potentially creating unrealistic instances), we will address this issue in the next section by performing perturbations that takes into considerations semantic relationships between features.

- The introduced approach is agnostic with respect to the black box because it does not use any internal model parameters. However, the current implementation of the presented methodology is only suitable for tabular data. In the next section, we will build on the above insights to develop a new method also applicable to sequential data.

5.4 Doctor XAI: sequential and ontology-linked data

In this section, we introduce **Doctor XAI**, a novel explainability technique able to deal with multi-labeled, sequential, ontology-linked data. *Doctor XAI* is a post-hoc interpretability method that focuses on *local* explanations, i.e., it explains the rationale behind the classification of a single data point. It is also *model-agnostic*, as it produces explanations whose computation is not based on the black-box inner parameters or structure. In this regard, *Doctor XAI* is similar to other black-box-agnostic techniques [302, 303, 143, 268]. However, to the best of our knowledge, ours is the first agnostic XAI technique applicable to sequential and ontology-linked data classification.

Given a patient whose clinical history classification needs an explanation, *Doctor XAI* first generates a local synthetic neighborhood around the selected patient exploiting the semantic information encoded in the ontology and uses the black-box model to label it. Then it transforms the clinical history of such synthetic patients into a format suitable to train a decision tree. This transformation allows taking the sequential nature of the data into account. Finally, *Doctor XAI* trains a decision tree on the labeled synthetic neighborhood, and it extracts an explanation in the form of a decision rule.

We applied *Doctor XAI* to explain the decisions of *Doctor AI* [72], a recurrent neural network which takes as input patients' sequential EHR data and predicts the next visit set of diagnoses. We compared the quality of the explanations provided by *Doctor XAI* against those of the same technique without the ontological information. We show how exploiting the semantic information encoded in the ontology increases the performance of the explainability technique across all the evaluated metrics. We want to highlight that, even if our system deals by design with sequential, multi-labeled, ontology-based data, none of these features is strictly necessary: *Doctor XAI* can be used with datasets displaying any combination of the three aforementioned features, by exploiting only the corresponding specific modules.

5.4.1 Ontology use in machine learning and XAI

In our work, we exploit the ontology of ICD-9 diagnosis codes (see section 5.1.5) to increase the fidelity performance of the interpretable model to the black-box. The increase in predictive performance, thanks to the infusion of knowledge in the learning procedure, was adopted in several other works. For example, in [73], the authors use an attention mechanism that leverages the medical ontology of ICD-9 to learn a code representation that combines the embeddings of its ontology ancestors. They then train this attention mechanism together with an RNN with GRU units to improve the classification performance of prediction of the predictive model. They show that the performance is increased by 10% with respect to a basic model that does not exploit the medical ontology. Furthermore, they show that the learned representation of medical codes aligns with medical knowledge. Moreover, the authors of [283] show how disease classification performance can improve using features based on the ICD-9 codes semantic similarity. To compute the ontological similarity among sets of ICD-9 codes, i.e., a visit, they first calculate the semantic similarity of each pair of terms in the sequences as the *importance* of their lowest common ancestor in the hierarchy and then take the maximum of these similarities as the similarity of the two sequences. This approach over-estimates the similarity of the two sequences since it is sufficient to have one ICD-9 code in common to have similarity equal to one. The *importance* of the lowest common ancestor is related to the level of the term in the hierarchy; according to the authors this feature is related to the rarity of the disease, but it just captures how well specified is the disease. However, even with this basic approach to encoding medical knowledge into the learning process, the performance of the algorithms is increased. We use a more sophisticated approach to compute patients similarity as detailed in Section 5.4.3. Closely related to ours, is the work presented in [82] where the authors use ontologies in the training of the surrogate model. In particular, they used a custom ontology to constrain the training of a surrogate global decision tree (DT) and perform a user study proving that if the nodes of the DT represents more general concept, the understandability of the explanation increases.

5.4.2 Methods

In this section, we introduce the components of *Doctor XAI* and how they form the full explanation pipeline. Our technique solve the *outcome explanation problem* (detailed in section 5.1.1) by learning an interpretable classifier able to mimic the decision boundary of the black-box that is relevant to the decision taken for a particular instance. In other words, given an instance x and its black-box outcome $y = b(x)$, an explanation is extracted for this individual decision from an inherently interpretable model c trained to mimic the local behavior of b.

For our approach, we follow the pipeline already presented in section 5.3 of generating a set of synthetic instances (the *synthetic neighborhood*) surrounding the instance x we want to explain by perturbing a set of *real neighbors* taken from a set of available instances, then labeling them utilizing the black-box b, training an interpretable model c on such labeled neighborhood, and finally extracting an explanation in the form of a symbolic rule. However, we have developed specific modules in order to deal with the temporal dimension in the data and exploit linked structural knowledge representation: Figure 5-7 illustrates our explanation pipeline.

5.4.3 The explanation pipeline

The starting point is the data point whose black-box prediction we are interested in explaining. As the first step, we select the data points that are closest to the instance to be explained in the available dataset: these points are called the **real neighbors** of the instance.

We can either select the closest data points according to a standard distance metric, such as the Jaccard one or exploit ontology-base similarities. We describe the latter in Subsection 5.4.3. In both cases, we obtain a set of real neighbors, each of which is represented as a sequence.

We then generate the **synthetic neighborhood** perturbing the first real neigh-

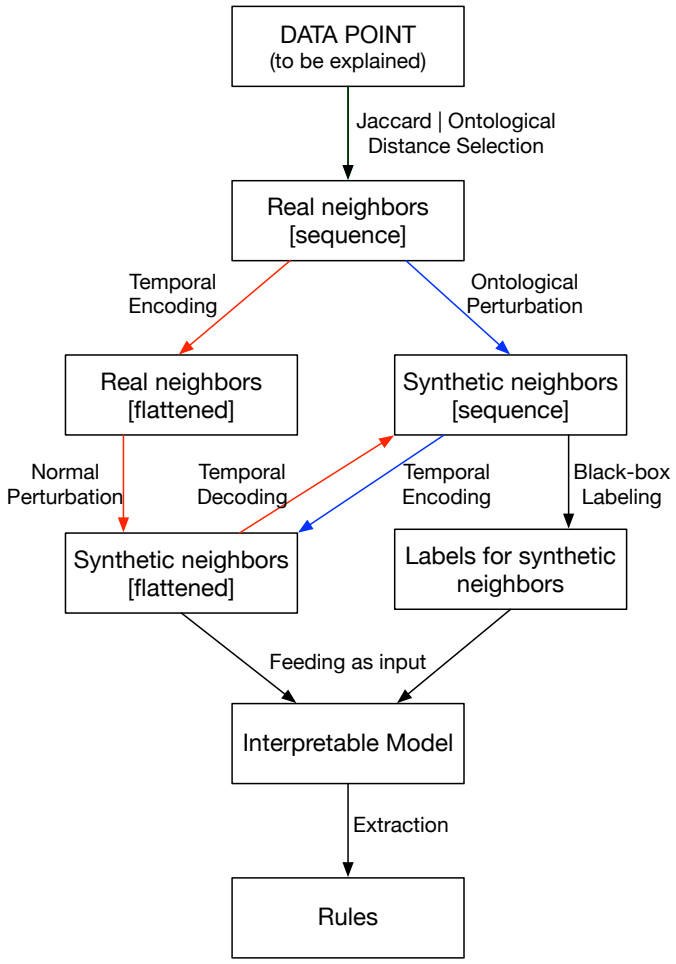


Figure 5-7: The explanation pipeline

bors to ensure the locality of the augmented neighborhood. The synthetic neighbors' sampling is crucial to the purpose of auditing black-box models.

Ideally, the synthetic instances should be drawn from the true underlying local distribution. Unfortunately, this distribution is generally unknown, and how to generate meaningful synthetic patients is still an open question. While most state-of-the-art agnostic explainers employ random perturbations, we use the domain knowledge encoded in the ICD-9 ontology to generate more meaningful synthetic instances, as explained in Subsection 5.4.3. It could be argued that the interpretable model could be trained directly on the closest real neighbors. However, the rationale behind the generation of synthetic neighbors is that we want to build a dense training set for the interpretable classifier c in order to increase its performance in

mimicking the black-box.

Unlike other explanation techniques, we do not perturb directly the features of the instance whose black-box decision we want to explain. By doing so, we prevent the case of generating a synthetic neighborhood containing only instances with the same black-box classification - a situation that would make the training of any interpretable model impossible. In other words, we ensure the *expressiveness* of the synthetic neighborhood, i.e., the black-box classifications are heterogeneous among the synthetic neighbors.

For the **perturbation** steps in our pipeline, we can follow two alternative paths, represented by the red and blue arrows in Figure 5-7 (the two paths share the black arrows). The red path is based on the **normal perturbation**, which we describe in Subsection 5.4.3; the blue path involves the **ontological perturbation**, as described in Subsection 5.4.3. Both paths involve steps of temporal encoding/decoding (with the relative algorithms described in Subsection 5.4.3), since the black-box model requires a sequential input, whereas the interpretable one requires a tabular (flat) one.

The red path is based on the **normal perturbation**: first, the real neighbors are encoded (flattened) into sparse vectors. Then the normal perturbation is applied in order to obtain a synthetic neighborhood - and this kind of data can be fed to an interpretable model. In order to obtain the labels for the synthetic data points, however, we have to decode them (back into sequences) so that we can feed them to our black-box model for labeling. Once we have both the synthetic neighborhood and the corresponding labels, we can train the interpretable model, and finally, extract symbolic rules.

Similarly to [268], we chose a multi-label decision tree as inherently interpretable classifier c. From such decision tree, we extract rule-based explanations in the form $p \rightarrow y$ where $y = c(x)$. The explanations are extracted by including in the rule

premise p all the split conditions on the path from the root to the leaf node that is satisfied by the instance x.

The blue path involves the **ontological perturbation**. In this case, we can apply the perturbation directly on sequential data, obtain a synthetic neighborhood as a set of sequences, and feed them to the black-box model for labeling. However, as it was for the red path, the interpretable model requires a tabular input, so we proceed to flatten (time-encode) the synthetic neighbors in a set of vectors. At this point, the blue path follows the same final steps as described above: training of the interpretable model and extraction of symbolic rules.

We remark that, while we followed a general framework for our model-agnostic explanation pipeline, we have extended the framework with novel contributions in order to deal with structured data and sequential data respectively. We observe that these components can be independently plugged in an explanation pipeline according to the nature of the data point to be explained.

Ontological Neighborhood

In this section, we define a new distance measure that allows us to select the semantically closest neighbors of the instance whose decision we want to explain.

Each patient's clinical history is represented as a list of visits, which in turn are encoded as lists of ICD-9 codes. Every instance is therefore a list of lists of ICD-9 codes. More formally, if we define the set of ICD-9 codes as $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$, each patient's clinical history is represented by a sequence of visits $V_1, \ldots, V_M$ such that $V_i \subseteq \mathcal{C}$. A simple example of a patient clinical history representation is as follows:

$$[[433.10, 453.81], [453.81], [453.81, 788.5, 790.01]]$$

The patient visited the hospital three times; the condition 453.81 (Acute embolism and thrombosis of superficial veins of unspecified upper extremity) is chronic, con-

dition 433.10 (Occlusion and stenosis of carotid artery without mention of cerebral infarction) was observed on the first visit only, whereas two new conditions (with codes 788.5 and 790.01) were diagnosed only in the third visit.

We observe that multi-hot encoding all occurring ICD-9 codes is a fairly inefficient representation for visits - the obvious drawback being the size of the encoding vector corresponding to the size of the ICD-9 dictionary. Furthermore, this positional representation does not encode the semantic distance from ICD-9 codes - a patient with food poisoning, one with a broken hand and one with a broken wrist are equally distant from a purely Hamming-based perspective. In order to mine the semantically similar data points, we introduce an ontology-based distance metric.

Code-to-code similarity Each ICD-9 code represents a medical concept in a hierarchical ontology, these concepts are the nodes of the graph-representation of such ontology, and it is therefore possible to compute distance and similarity scores among any pair of them. Several similarity metrics could be selected; in this paper, we adopt the Wu-Palmer similarity score (WuP) [389] because it is one of the most commonly used for ICD-9 ontologies [180, 14, 131]. Given two ICD-9 nodes c_1 and c_2, let L be their lowest common ancestor (LCA) and R be the root of the ICD-9 ontology; also let $d(x, y)$ be the number of hops (steps) required to reach node y from node x following the ontology links. The WuP similarity measure between c_1 and c_2 corresponds to:

$$WuP(c_1, c_2) = \frac{2 * d(L, R)}{d(c_1, L) + d(c_2, L) + 2 * d(L, R)}$$

$WuP(c_1, c_2) \in [0, 1]$ for any couple of ICD-9 nodes. The lower bound 0 is obtained when $d(L, R) = 0$, that is, when the LCA of c_1 and c_2 is the root node. Conversely, a node has WuP-similarity 1 with itself. By relying on the underlying ICD-9 ontology, we can therefore use the WuP similarity to compute pairwise distances between ICD-9 codes. This yields a much more fine-grained analysis compared to a coarse Hamming similarity.

Visit-to-visit distance Having defined a code-to-code distance, the following step is to compute distances at the visit level - since visits are defined as lists of occurring ICD-9 codes. We adopted the weighted Levenshtein [213] distance, a string metric for measuring the difference between two sequences as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one sequence into the other. The weighted version of the Levenshtein distance allows defining custom insertion/deletion/edit costs. We have set $1 - WuP(c_1, c_2)$ as edit cost for modifying c_1 into c_2, and 1 as insertion/deletion (indel) cost (since $WuP(c_1, c_2) \geq 0$, $1 - WUP(c_1, c_2) \leq 1$) in order to favor edits over indels. This gives us a distance metric between pairs of visits which is based on the similarity between the ICD-9 codes occurring in each of the two visits.

Patient-to-patient distance The third step is to compute a patient-to-patient distance metric based on how similar the visits of the two patients are. In order to do so, we adopted the Dynamic Time Warping (DTW) algorithm [38], again using the pairwise visit distances provided by the weighted Levenshtein algorithm as edit distance. The sequences of visits are *warped* non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This final step provides us with the pairwise distances for all patients (data points) in the dataset, thus enabling us to select real neighbors with ontologically similar conditions w.r.t. the data point to explain.

Ontological Perturbation

As previously mentioned, after selecting the first real neighbors of the instance whose decision we want to explain, we perturb them in order to generate synthetic neighbors. There are mainly two ways to perform an ontology-based perturbation on an instance: by masking or replacing some conditions (ICD-9 codes) in the patient's clinical history according to their relationships in the ontology. We decided to adopt the first type of perturbation in order to limit the amount of noise injected in the training set of the interpretable classifier. The idea behind perturbing the patient's

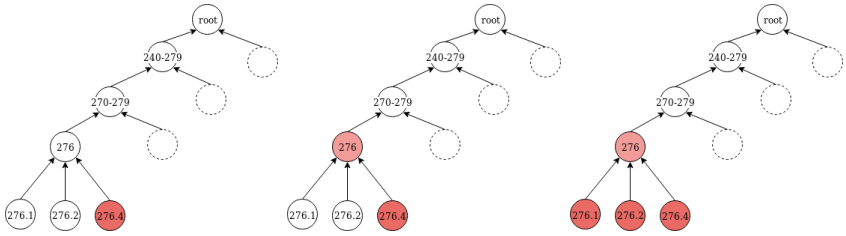


Figure 5-8: *(1st plot)* The node corresponding to the randomly selected ICD-9 code (276.4) of the patient is highlighted in red in the ICD-9 ontology graph representation. *(2nd plot).* The ontological superconcept of the selected ICD-9 is selected and highlighted (276). *(3rd plot)* All ICD-9 codes all having as parent the identified superconcept are selected and removed from the patient (codes 276.1, 276.2 and 276.4).

history in this way is that we want to explore how the black-box label changes if we mask all the semantically-similar items from the sequence. Furthermore, the ontological perturbation of instances takes into account by-design the relationships among the single features (in this case the ICD-9 codes) thus creating more realistic synthetic instances. We decided to randomly mask all the occurrences of the items with the same least common superconcept. By doing so, we are exploring how a general condition (a higher concept in the ontology) is affecting the black-box diagnosis. In our case, we are dealing with patients' clinical history. Each patient's clinical history is a sequence of visits, and each visit is represented by lists of ICD-9 codes. In the ICD-9 ontology, all codes are composed of a prefix and a suffix, separated by a dot: the prefix defines the general condition, and the suffix provides increasingly specific information. We show an example of the hierarchical structure of the ICD-9 ontology in Figure 5-2. Our implementation of the ontological perturbation is the following: We first randomly select one ICD-9 code in the clinical history of the patient we want to perturb (a leaf of the ontology), then we mask all the ICD-9 codes in the patient's history that share the same prefix (the least common superconcept). By doing so, we generate synthetic patients that lack a specific group of semantically similar conditions.

Consider, for example, the following patient:

$$P = [[276.1, 276.2], [276.4, 530.1], [507, 530], [276.2, 530.19]]$$

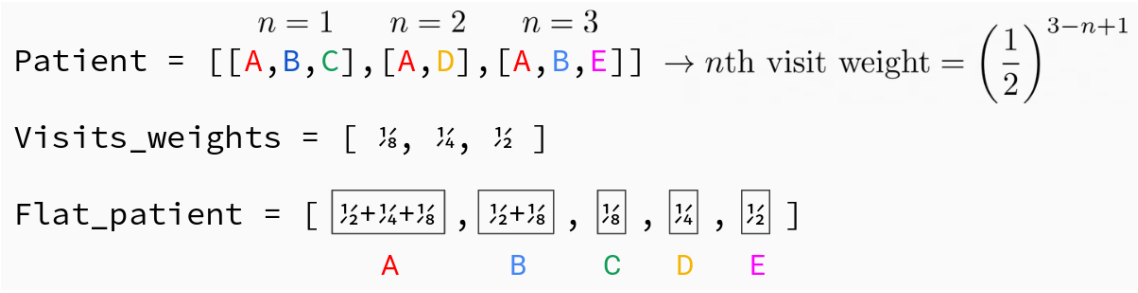


Figure 5-9: Example of temporal encoding for a patient

One example of ontological perturbation is the following: we randomly select ICD-9 code 276.4 which is *mixed acid-base balance disorder*. Starting from this code we create the synthetic patient

$$P^* = [[], [530.1], [507, 530], [530.19]]$$

by masking all the ICD-9 codes related to ICD-9 276, i.e., *disorders of fluid electrolyte and acid-base balance* (the least common superconcept). A graphical representation is shown in Figure 5-8. Note that, without ontological information, we have 7 different codes and therefore 2^7 potential perturbations, most of which don't really isolate different conditions. Conversely, using the ontology we group the occurring ICD-9 codes in three categories {276*, 507*, 530*}: as a consequence we have 8 potential maskings, each of which isolates a subset of different conditions.

Normal Perturbation

As an alternative to the ontological perturbation of the first real neighbors of the instance under study, we performed a *normal perturbation* on such features. This perturbation applies to a broader number of cases since it does not require an ontology to be performed. Given the *flattened* version of the real neighbors, the normal perturbation creates the new synthetic instances feature by feature drawing from a normal distribution with mean and standard deviation of the empirical distribution of that feature in the real neighbors. This perturbation implies the strong assumption that every feature is independent of the others.

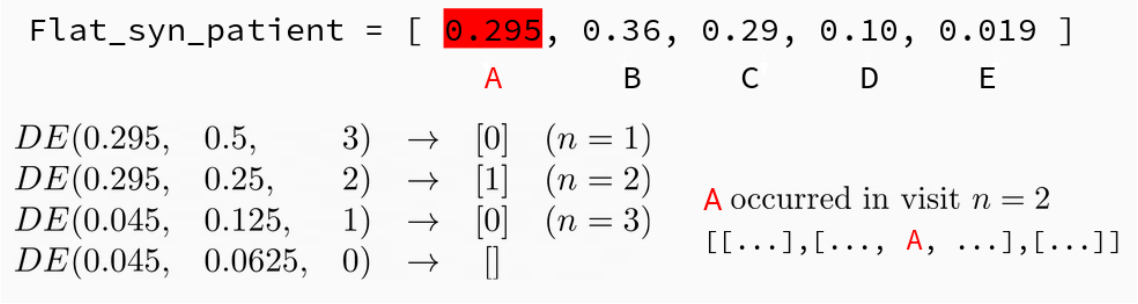


Figure 5-10: Example of temporal decoding for a patient

Temporal encoding and decoding

As introduced above, the standard data type for longitudinal healthcare data is to represent a patient as a list of visits, and in turn each visit as a list of occurring conditions (in our case, ICD-9 codes). There is no inherently interpretable model able to deal with the multi-label classification of such type of input; therefore, we need to perform an input transformation that both retains its sequential information and allows to feed it into an interpretable model - a decision tree in our case.

We introduce a pair of encoding-decoding algorithms so that we can *flatten* the temporal dimension when feeding our synthetic neighborhood to the interpretable model. The binary encoder implements a time-based exponential decay rooted at the last item of the sequence. Intuitively, each code c_i in visit V_j will be given a score of $+.5$ if V_j is the last visit, $+.25$ if V_j is the second-to-last visit, and so on. More formally, when encoding a patient $P = [V_1, .., V_N]$, each code $c \in P$ will be encoded as follows:

$$EN(c, P) = \sum_{i=1}^{n} (1/2^{n-i+1} \text{ if } c \in V_i \text{ else } 0)$$

The encoding is 0 for all items that never occur in that sequence, and it tends to 1 for a growing number of elements in the sequence in which that item occurs. The encoded (flattened) representation of a patient is therefore a sparse vector of real numbers, and as such it can be fed to multiple interpretable models.
Conversely, we define the decoding from a sparse vector of real numbers to a sequence

of visits as:

$$DE(X,t,l) = \begin{cases} [] & \text{if } X = 0 \text{ or } l = 0 \\ append(DE(X-t, t/2, l-1), [1]) & \text{if } X > t \\ append(DE(X, t/2, l-1), [0]) & \text{otherwise} \end{cases}$$

where X is the value to be decoded, t is initially set at .5 and l controls the maximum length of the generated sequence (we use the average length of the real neighbors). The result of the decoding is a list of 0s and 1s that indicates the presence/absence of a certain code.

We show a simple example of temporal encoding in Figure 5-9. In this example, the patient visited the hospital three times. Each visit contains a set of ICD-9 codes (for the sake of simplicity here represented as letters). As a first step, a weight is associated to each visit. Then the weight of each ICD-9 code is computed by adding the weights of the visits where it occurred. We also show a simple example of temporal decoding of a flat synthetic patient in Figure 5-10. In this example, we transform the value of the first ICD-9 code (represented by letter A) into its occurrence in the sequence. In this example we set the maximum length of the generated sequence to $l = 3$. It is important to remark that the decoding algorithm, when presented with perturbed data, might potentially produce arbitrarily long sequences, where progressively small residuals are mapped to the occurrence of the decoded ICD-9 code in progressively further away visits. The l-guard was introduced to prevent this from happening so that flattened synthetic patients match the number of visits of the flattened real neighbors.

5.4.4 Experiments

Dataset

We ran our experiments on the *Multiparameter Intelligent Monitoring in Intensive Care III* (MIMIC-III) database [188]. This database contains de-identified data of over 40.000 ICU (Intensive Care Unit) patients of the Beth Israel Deaconess Medical

Center data in Boston collected from 2001 to 2012. We used the information related to the hospital stay (dates and diagnosis codes) to build the patient clinical history as performed by the pre-processing script available in *Doctor AI* GitHub repository[6]. This operation removes all patients with less than two visits, some statistics about the dataset after the pre-processing procedure can be found in Table 5.9.

	MIMIC-III
n. of patients	7499
n. of visits	19911
avg. n. of visits per patient	2.65
min. n. of visits per patient	2
max. n. of visits per patient	42
n. of unique ICD-9 codes	4880
n. of unique CSS grouper codes	272
avg. n. of ICD-9 codes per visit	13.06

Table 5.9: MIMIC-III characteristics for patients with more than one visit

The clinical history of each patient is modeled as time-stamped sequence of visits. As previously mentioned, each visit is represented by a set of ICD-9 diagnosis codes, these codes are assigned to each patient at the end of his or her hospital stay, and hospitals use them to bill for care provided. They are organized in a "is-a" hierarchical tree structure[7] that places more general concepts closer to the root of the tree and more fine-grained concepts closer to the leaves of the tree. The ICD-9 taxonomy and occurring ICD-9 codes in MIMIC are visualized in Figure 5-11. We used this ontology to measure the similarity between patients' clinical history as described in section 5.4.3 and to generate the synthetic neighbors of each patient as described in section 5.4.3.

Black-box classifier

Doctor AI [72] is a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU) that predicts the patient's next visit time, diagnoses and medications order. We focus here only on the diagnosis prediction task of the model. The authors trained their model on 260.000 patients of the EHRs database of Sutter Health Palo

[6]https://github.com/mp2893/doctorai
[7]https://bioportal.bioontology.org/ontologies/ICD9CM

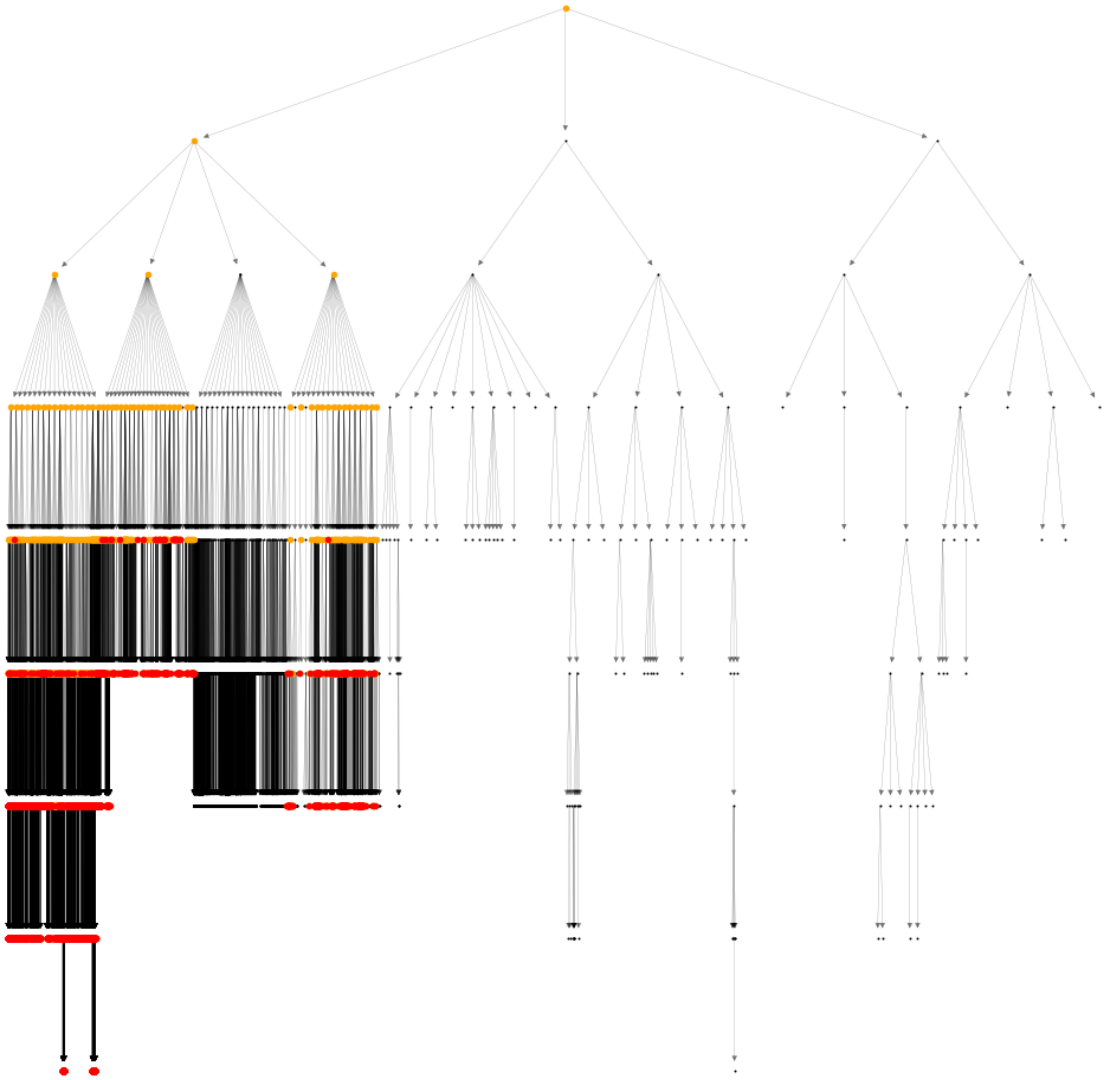


Figure 5-11: ICD-9 ontology. The red dots represent codes occurring in the MIMIC dataset, the orange ones their parent nodes.

Alto Medical Foundation. The multi-hot input vector representing the diagnoses at each time-step of patient clinical history is first projected in a lower-dimensional space and then received as input by a stack of RNN layers implemented using GRUs. Finally, a Softmax layer is used to predict the diagnosis codes of the next time-stamp. The predictive performance of *Doctor AI* are evaluated using recall@n with $n = 10, 20, 30$ achieving 0.79 recall@30.

We trained *Doctor AI* on *MIMIC-III* for 50 epochs, using approximately 70% of patients as the training set, 15% as the validation and 15% as the test set.

Table 5.10: *Doctor AI* performance on different datasets.

Dataset and algorithm	recall@n		
	n=10	n=20	n=30
Doctor AI: MIMIC-III	0.350	0.521	0.631
Most frequent: MIMIC-III	0.383	0.473	0.491
Doctor AI: dataset from [72]	0.643	0.743	0.796
Most frequent: dataset from [72]	0.566	0.674	0.717

We built the label for each time step of the sequence by grouping the full-length ICD-9 codes using CCS single-digit groupers[8]. By doing so, the dimensionality of the label space shrinks from 4880 codes to 272 groups of codes. We compare the predictive performance of *Doctor AI* trained by us on MIMIC-III dataset with the ones reported in the original paper in Table 5.10. We also trained a baseline model to imitate one of the benchmarks of the original paper. This baseline, the *Most frequent*, predicts the top-k most frequent labels observed in visits before the current visit. The fact that we trained *Doctor AI* on a much smaller dataset lowers the algorithm's predictive performance compared to the ones of the original paper. However, they are in line with the performance on the MIMIC-II dataset discussed in the original paper. Furthermore, having a good predictive performance is not our goal; we will use the black-box labels as ground-truth labels for the decision tree. In our work, we focus on explaining *Doctor AI* because of the availability of its source code and because the authors' results are easily reproducible using open-source data. However, we want to stress that our method is not specific to this black-box.

Experimental set-up

We decided to test our explanation method on a cohort of 1.000 randomly selected patients from the MIMIC database. We put each of these 1.000 patients through 3 different explanation pipelines and we explained their top-10 CCS-codes prediction. The first two exploit the ontological information encoded into ICD-9 codes, whereas the last one can also be used to explain sequential data classification if an ontology is missing. We aim to show that exploiting the ontological information in the data

[8]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

increases the explanation quality.

- *Ontological pipeline with ontological perturbation - Dr.XAI.* This pipeline fully exploits the knowledge encoded into the ICD-9 ontology to create the synthetic neighborhood. Given a patient whose black-box decision we want to explain, it selects its first k neighbors in the dataset using the *ontological distance* described in section 5.4.3 and then it generates the synthetic neighborhood by perturbing them using the *ontological perturbations* described in section 5.4.3. This pipeline corresponds to the blue path of Figure 5-7 using the Ontological similarity.

- *Ontological pipeline with normal perturbation.* This pipeline selects the first k real neighbors of the instance to explain using the *ontological distance*, but then it creates the synthetic neighborhood by perturbing these instances using the *normal perturbation* described in section 5.4.3. This pipeline corresponds to the red path of Figure 5-7 using the Ontological similarity.

- *Non-ontological pipeline with normal perturbation.* This pipeline does not use the semantic information encoded in the ICD-9 codes. It first selects the k real neighbors of the instance to be explained using *Jaccard similarity* between each patient visit and then it perturbs them by using *normal perturbations* 5.4.3. This pipeline corresponds to the red path of Figure 5-7 using the Jaccard similarity.

By comparing the two ontological pipelines, we want to show that exploiting the semantic information encoded in the ICD-9 ontology is also useful to create the synthetic neighbors. We developed the *non-ontological pipeline* as a baseline for explanation quality. However, this last pipeline is also the most general one because it can be applied to sequential data that does not have an associated ontology. Furthermore, we wanted to show that increasing the density of the feature space around the instance to be explained by creating the synthetic neighbors actually increases the interpretable model's ability to mimic the black-box locally. For this reason, for each instance to be explained, we trained two decision trees. One decision tree is trained directly on the real neighbors of that patient from the dataset, and

the other one is trained on a fraction of the augmented synthetic neighborhood. We then compare the performance of these decision trees on an out-of-sample set of synthetic neighbors. We utilize the following metrics to evaluate and compare the different explanation pipelines.

- *Fidelity to the black-box* $\in [0, 1]$ This metric compares the predictions made by the interpretable model with the predictions made by the black-box on a synthetic neighborhood of the instance. It measures the ability of the interpretable classifier to locally mimic the black-box, and therefore it is tested on a held-out subset of the synthetic neighborhood. Since we are dealing with a multi-label classification task, we calculate the fidelity the F_1 measure with micro-averaging [406].

- *Hit* $\in [0, 1]$ This metric compares the interpretable classifier prediction y_c and the black-box prediction y_b on the instance to be explained. It tells us if the interpretable classifier predicts the same label as the black-box on the instance we want to explain. Since the prediction we are trying to explain is a multi-label classification, we calculate the hit as $1 -$ hamming-distance(y_b, y_c).

- *Explanation complexity.* This metric measures the complexity of the explanation as the number of premises in the rule-based explanation. This measure is important since we do not want to approximate the black-box with a model that loses its interpretability because of the high-dimensionality of the explanations it produces [223, 98].

5.4.5 Results

In Figure 5-12 we show the fidelity sample distributions at different values of k for the decision trees trained using the *ontological explanation pipelines*, i.e., the pipelines that select the first k dataset neighbors of the instance to be explained using the *ontological distance*. The first observation is that the decision trees trained directly on the k real neighbors (blue and green boxplots) generally have a lower fidelity to the black-box compared to the ones trained on the augmented synthetic neighborhood (orange and red boxplots). This trend is true for all values of k and for both the

ontological pipeline with ontological perturbation and the *ontological pipeline with normal perturbation*. The fidelity values of each decision tree have been evaluated on an held-out test set of synthetic neighbors. This trend confirms that increasing the local density of points in the feature space around the instance to be explained helps the interpretable model to understand the black-box behavior. The second observation is that the fidelity of the decision tree trained using the *ontological pipeline with ontological perturbation* (red boxplot) is generally higher compared to all the other explanation pipelines. This observed tendency confirms that exploiting the ontological information during the synthetic neighborhood creation allows the decision tree to better approximate the local black-box decision boundary.

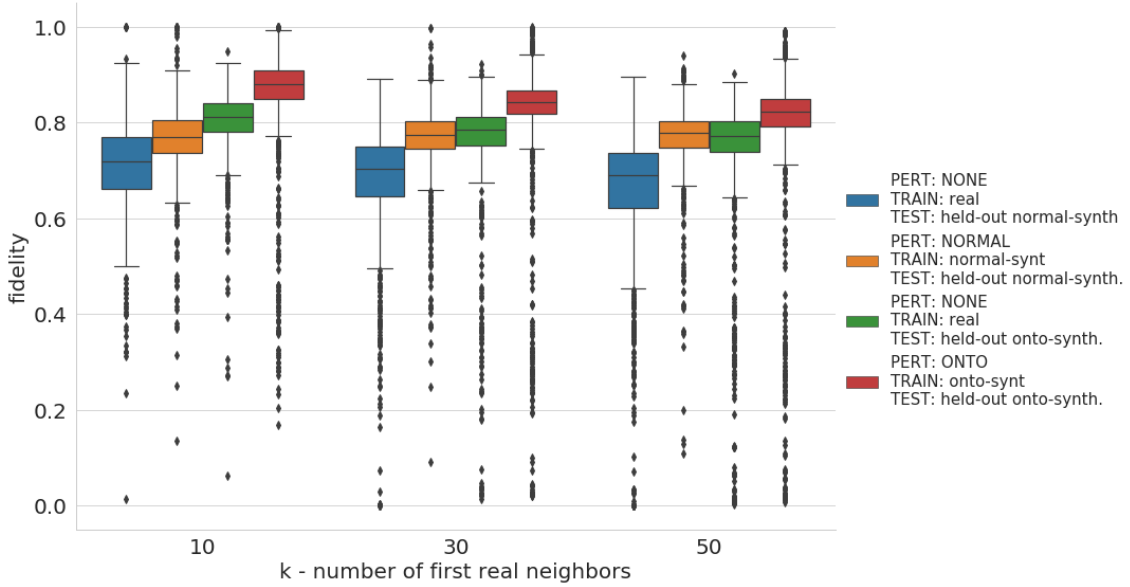


Figure 5-12: Fidelity distribution for the ontological pipeline with different k, perturbation type, and training/test set.

In Figure 5-13 we show the fidelity sample distributions at different values of k for the decision trees trained using the *non-ontological explanation pipeline*, i.e., the pipeline that selects the first k dataset neighbors of the instance to be explained using the *Jaccard similarity* between patients' visits. We developed this explanation pipeline that does not use the semantic information encoded into the ICD-9 codes as a baseline to prove that an approach that does not exploit this information has lower performance. This is true if we compare this explanation pipeline with the fully-ontological one (the *ontological pipeline with ontological perturbation*). However, the

fidelity performance of this non-ontological pipeline is comparable to the ones of the *ontological pipeline with normal perturbation*. The high values of fidelity achieved by this pipeline prove that we developed a *trustable* explainability technique applicable to any black-box that takes as input any sequential data, even when there is no ontology associated with the items of the sequence. Furthermore, it is important to notice that, also for this pipeline, the values of fidelity to the black-box increase after the synthetic neighborhood augmentation (the orange boxplot).

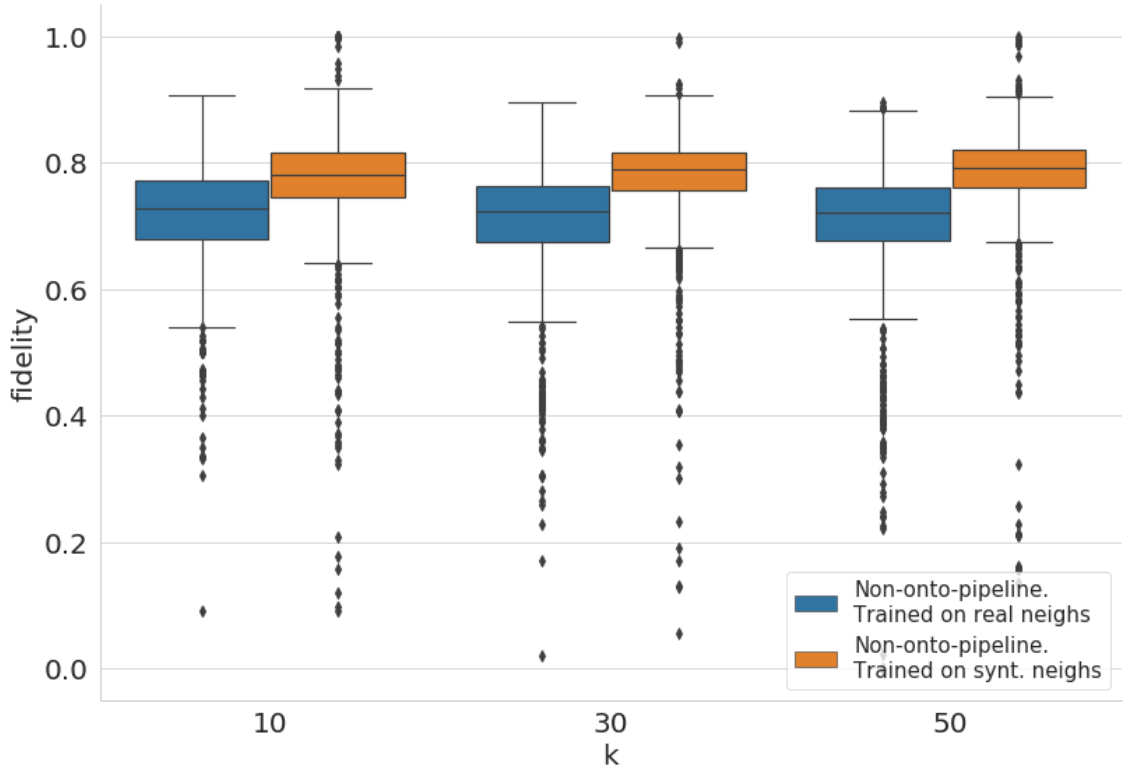


Figure 5-13: Fidelity distribution for the non-ontological pipeline at different values of k and training set.

In Figure 5-14 we show the sample distribution of *explanation complexity*, i.e., the number of premises in the rule-based explanations at different values of k for the two ontological explanation pipelines. As expected, we see how the length of the explanation increases as k increases. This happens because if we start from a high number of first real dataset neighbors we are trying to approximate a larger portion of the decision boundary of the black-box with the interpretable classifier. We could say that we are not restricting ourselves to the *local* decision boundary close to the instance whose decision we want to explain. Therefore, since we are trying

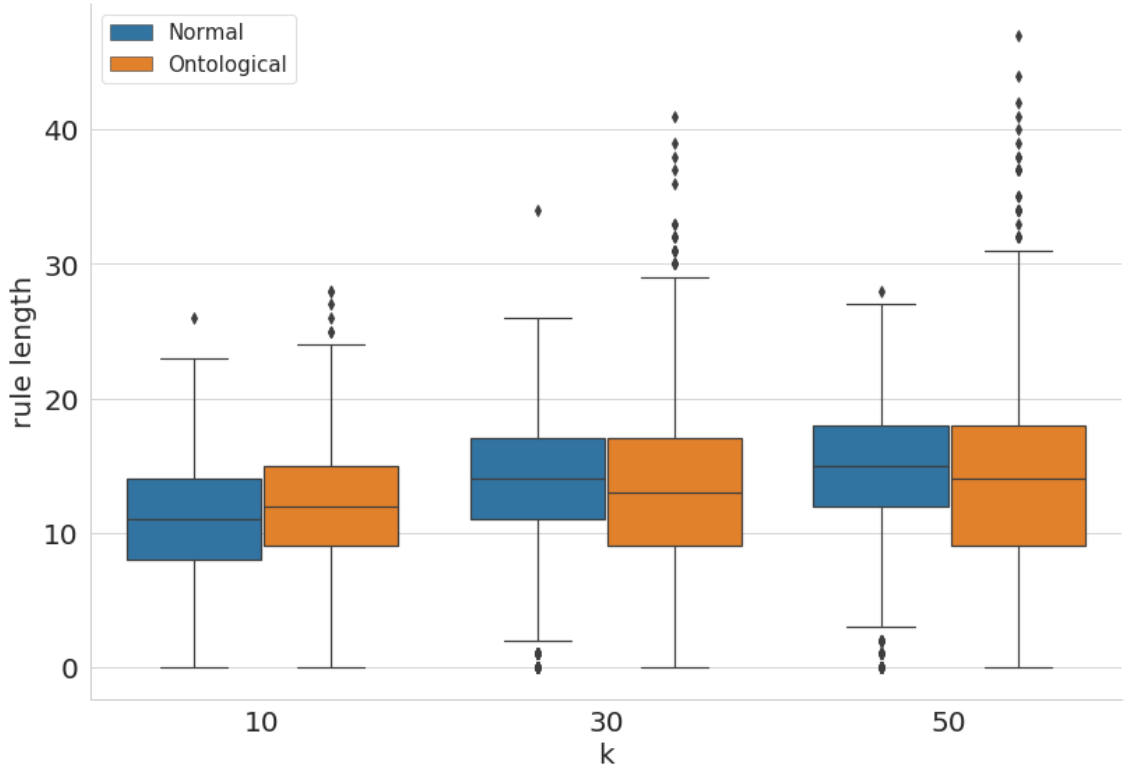


Figure 5-14: Explanation complexity for the ontological pipelines

to approximate a more complex decision boundary the dimensionality/complexity of the decision tree grows and consequentially the length of the rule increases. From this plot it is also possible to see that the explanation length of the explanations extracted from the *ontological pipeline with ontological perturbation* (orange boxplot) is more variable than the ones extracted using the *ontological pipeline with normal perturbation* for large values of k.

Aggregated statistics of fidelity and of hit for all the explanation pipelines are shown in Tables 5.11 and 5.12: we can observe that the value of hit is consistently high for all explanation pipelines and across all values of k.

5.4.6 Explanation example

We show in Figure 5-15 an explanation example extracted with the *ontological pipeline with ontological perturbation* with $k = 10$. In order to make it more comprehensible for readers not familiar with ICD-9 codes, we enriched the rule-based

Table 5.11: Mean values of fidelity

Explanation Pipeline	Fidelity					
	k=10		k=30		k=50	
	realDT	syntDT	realDT	syntDT	realDT	synDT
Ontological pipeline with ontological perturbation	0.81	0.89	0.77	0.85	0.12	0.79
Ontological pipeline with normal perturbation	0.70	0.73	0.67	0.62	0.10	0.76
Non-ontological pipeline with normal perturbation	0.71	0.77	0.69	0.47	0.68	0.78

Table 5.12: Mean values of hit

Explanation Pipeline	Hit					
	k=10		k=30		k=50	
	realDT	syntDT	realDT	syntDT	realDT	synDT
Ontological pipeline with ontological perturbation	1.00	1.00	1.00	1.00	0.93	1.00
Ontological pipeline with normal perturbation	1.00	0.98	1.00	0.99	0.93	0.98
Non-ontological pipeline with normal perturbation	1.00	0.99	1.00	0.99	1.00	0.99

explanation with the ICD-9 codes semantic. The original decision rule extracted from the decision tree can be seen at the top of the figure with the fidelity of the decision tree and its hit value. There are several ways to read this rule since it contains many layers of information. The decision rule is the decision tree pathway that leads from the root of the tree to the leaf containing the black-box decision; for this reason, all inequalities are to be considered in conjunction - furthermore, the ICD-9 codes occurring in the rule are ranked in order of information gain. Each conjunct of the rule follows the pattern

$$\text{ICD-9\_code} = \textit{observed\_value} \gtrless \textit{threshold\_value}$$

The *observed value* is the value of that ICD-9 code for the patient whose decision we want to explain. Recall that the temporal encoding or *flattening* procedure described in Section 5.4.3 assigns to each ICD-9 code a weight according to the visit in which it was observed (diagnosed). The *threshold value* is the split value assigned by the decision tree to that ICD-9 code. Both these values can be interpreted as the presence of the ICD-9 code in a set of visits. The patient under examination had four visits. The ICD-9 codes describing the diagnoses associated with each visit are represented in the timeline just below the decision rule. Recall that we are explaining the top-10 CCS-codes predicted by *Doctor AI*. The ICD-9 codes considered meaningful by the black-box have been colored to enhance the readability. The explanation of each real and threshold value can be found in the list below the timeline. For example, the ICD-9 code 584.5 has an observed value of 0.25, which means that it was observed in the second-to-last visit (visit 3). Its threshold value is 0.12, whose closest value among those generated in the temporal encoding process is 0.125 which represents the third-to-last visit (visit 2). For this reason, even if this ICD-9 code was observed in the penultimate visit, the interpretation of the first rule conjunct is *584.5 has to have been observed at least once in the last three visits*.

The code to run our experiments as well as our results are available on GitHub[9].

[9]https://github.com/CeciPani/DrXAI

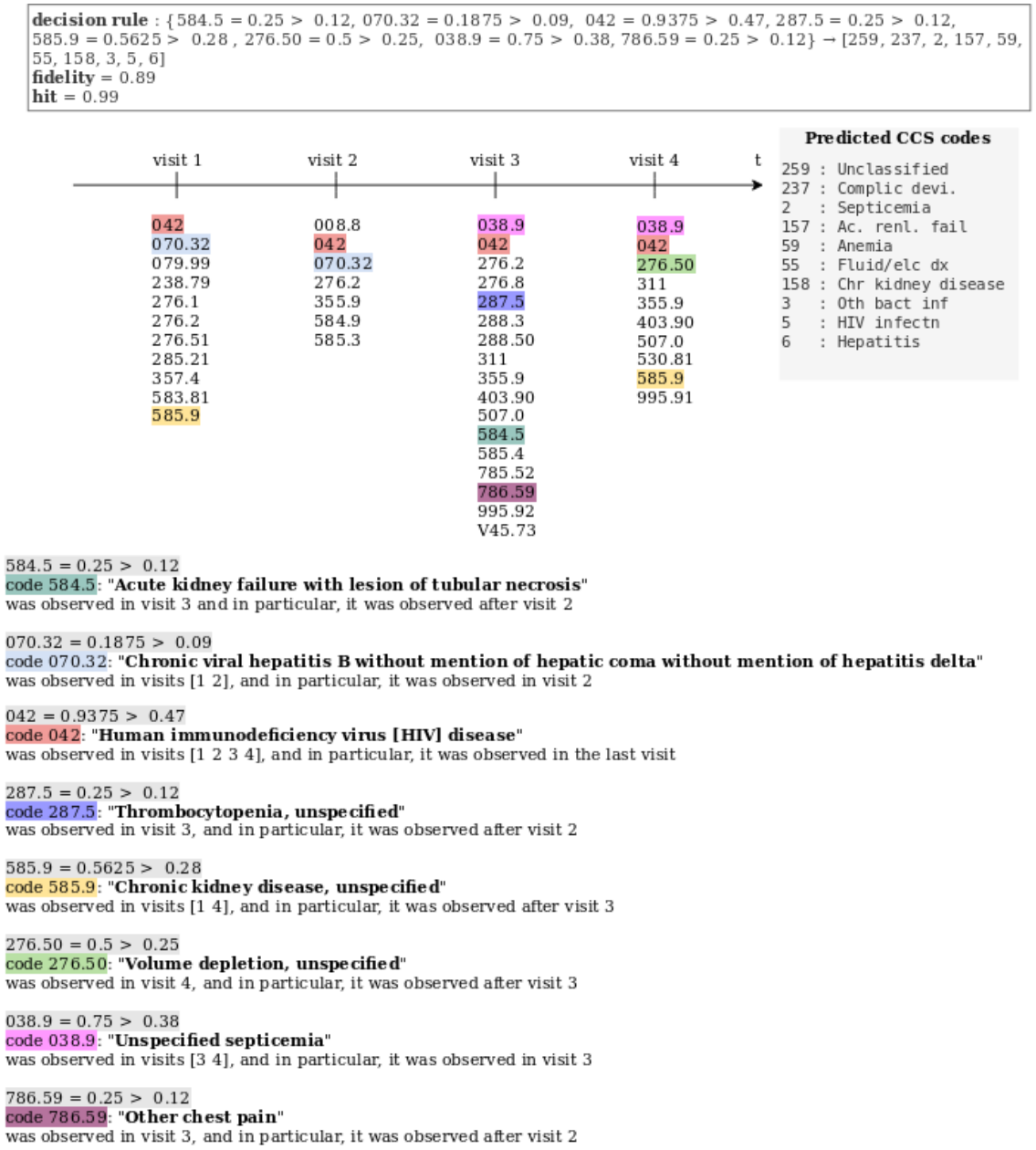


Figure 5-15: Explanation example

5.4.7 Lessons learned

We have presented **Doctor XAI**, the first model-agnostic approach to address the multi-label black-box outcome explanation problem for sequential and ontology-linked data. These features are typical of healthcare data. Our technique builds on the insights provided by the analysis of *MARLENA* (section 5.3). Indeed it employs the same pipeline: it first generates a set of synthetic instances close to the

instance whose black-box decision we need to explain, then it trains an interpretable classifier - a multi-label decision tree - on such neighborhood, and finally, it extracts a rule-based explanation from it. However, each step has been tailored to deal with sequential data and exploit ontology-linked data. In the following, a summary of some key aspects emerging from the analysis:

- Similar to the analysis conducted on *MARLENA* (section 5.3), we found that local multi-label decision trees can provide high-quality explanations in terms of fidelity to the black box, hit and compact rules.

- The synthetic augmentation of the interpretable classifier training set by perturbing a set of core real neighbors allows it to increase its fidelity to the black-box.

- Exploiting the semantic information encoded in medical ontologies in the creation of the neighborhood of the instance to be explained increases the quality of explanation in terms of fidelity, in particular:

  - If exploited in finding the set of real neighbors. We tested the sequential-only version of our explanation technique showing that it achieves good fidelity to the black-box, while also confirming that the ontology-enriched approach achieves a better score.

  - If exploited in the generation of synthetic neighbors. We studied the behavior of the interpretable classifier varying the hyper-parameter k - the number of first neighbors in the real dataset that are considered in the synthetic neighborhood generation. In particular, we showed that, for all values of k, the synthetic neighborhood generation procedure which exploits the ontological information encoded in the ICD-9 codes achieves better performance in approximating the local behavior of the black-box if compared to a procedure which does not have access to the ontology.

- The highest quality of explanations in terms of fidelity is probably due to the fact that the ontological perturbations used to create the synthetic neighbors take into account the relationships between features by design.

5.5 Discussion

In this chapter, we presented two novels XAI approaches to solve the outcome explanation problem for multi-label outcomes and sequential and ontology-linked data. This peculiar data characteristics are often found in healthcare data, as discussed in sections 5.1.3, 5.1.4 and 5.1.5. The presented approaches extracted an explanation for the black box decision on a specific data point x by first generating a synthetic neighborhood around such instance, then training a local surrogate model (in particular a decision tree) on this neighborhood, and finally extracting an explanation for the instance in the form of the decision path (decision rule) satisfied by x on the decision tree.

In section 5.3, we studied different ways to create an appropriate synthetic neighborhood for multi-label outcomes which is both *local* and *expressive* enough to capture the relevant part of the black box decision boundary. We created such synthetic neighbors by first selecting a set of *real* neighbors from a set of available instances and then perturbing them by sampling from their empirical local features distributions. We selected these *real neighbors* considering both distances in the features and labels space, following the hypothesis that this would have increased the local surrogate model fidelity to the black box. While this hypothesis turned out to be false, we learned that creating the synthetic neighbors starting from a set of k real neighbors and using multi-label decision trees allowed us to generate high-fidelity explanations. We also learned that the *local data density*, i.e., how close the *real* neighbors are to each other, highly impacts the fidelity of the explanations, making k an important hyperparameter of the explanation process.

In section 5.4, we presented an approach for the explanation of multi-label outcomes of black boxes trained on sequential ontology-linked data. This approach builds on the insights of section 5.3 by creating a synthetic neighborhood starting from a set of k *real* neighbors and training a multi-label decision tree to extract an explanation. However, we considered only neighbors in the feature space and we

tested our method for different values of k showing good fidelity performance across all tested values. We also showed that performing a synthetic local data augmentation by perturbing the k real neighbors always increased the local surrogate fidelity to the black box. Two key aspects of the presented approach are that it exploited the ontology in creating the synthetic neighborhood and employed a novel encoder/decoder scheme for sequential data that preserves the interpretability of the features. The ontological perturbation of the k real neighbors allowed to create synthetic instances that considered local features interactions. Indeed it perturbed the set of real neighbors by masking semantically similar features. This approach showed a great fidelity improvement with respect to a classical random perturbation approach.

Two types of criticisms are usually made on the type of approaches presented in this chapter, i.e., the post-hoc approach to explainability based on local surrogates. The first one is that they suffer from *explanation instability*, i.e., they might provide different explanations for the same instance in different iterations of the explainer [15, 146, 43, 372]. This is caused by the fact that the training set of the local surrogate model is stochastically generated by perturbing the available instances. The stability of the proposed approach was not the main focus of this thesis. However, it is a pertinent issue since explanation stability is a relevant requirement for high-stakes domains such as healthcare. Even though how to solve this problem is still an open question, the presented approaches could be optimized for explanation stability by creating multiple realizations of synthetic neighborhoods for the same instance and employ an appropriate aggregation technique to merge all the resulting explanations. However, this would highly increase the time needed to extract an explanation. Other solutions to this problem has been proposed by avoiding to generate a synthetic neighborhood [397] (however, as shown in this chapter, avoiding to generate a synthetic neighborhood has a cost in terms of the surrogate model fidelity) or by performing the perturbation process into a latent space [141, 327, 179]. Intuitively, if the number of dimensions of the latent space is correctly chosen, they should distill all the relevant information and variability of the local data distribution. However, when working in the latent space, it is difficult to understand each

latent feature's semantic meaning, which adds a layer of opacity to the explainability method. This opacity makes it difficult to determine the correct number of dimensions for the latent space. Furthermore, correctly training local embedding models such as autoencoders [141] could be taxing.

The second type of criticism is based on the fact that black box models have become the default choice even when not needed [306] because they are wrongly perceived as being more accurate than their interpretable counterparts. Therefore the research community should *"stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"*. This claim is based on the fact that Data Science is an iterative process that includes moving back and forth between the different stages of problem understanding, data preparation, modeling and evaluation. Furthermore, post-hoc explainers force the user to rely on the accuracy of two models instead of one and that they could be misleading in a number of different ways. While this is a valid criticism, we argue that post-hoc explainability techniques play an indispensable role for the development of AI applications and in the knowledge discovery process overall for many reasons.

First, they provide an additional and much-needed debugging tool for machine learning engineers [39]. The combined use of black box algorithms and post-hoc XAI techniques allows the developer to explore the data and model at hand that, in principle, could allow her to gain enough insight to build an interpretable model. Second, they allow a sanity check for sophisticated models that solve critical tasks when there is no alternative. Indeed there is no guarantee that the back and forth between the different phases of the data mining process makes it possible to develop an interpretable model with the same level of performance as a more complex one. Moreover, even if preliminary work on theoretical guarantees of the existence of such models exists [320], it might be challenging to find it in practice. Furthermore, the two lines of research on XAI and on Interpretable ML are not mutually exclusive. However, we agree that the field of XAI is still in its infancy, and it needs to solve a number of critical challenges before being ready to be deployed. Furthermore, when

developed correctly, *faithful*, *stable* and *trustworthy* post-hoc explanations could, in principle, reconstruct reasonings that differ considerably from the human one, but which allow learning something new. This would be especially interesting in AI applications such as pharmacology, where AI algorithms are employed to discover new drugs. Lastly, they are the only solution to auditing black boxes covered by industrial and commercial secrecy, i.e., commercial black boxes. A use case where the Doctor XAI 5.4 is used to perform external auditing of a black box clinical decision support system is presented in next chapter 6.

Furthermore, while in this chapter we focused our evaluation on technical metrics of explanation goodness such as fidelity and explanation complexity, in chapter 7, we study how an explanation could impact users trust in the AI system and their behavioral intention of use such systems.

Chapter 6

XAI to audit clinical decision support systems that are proprietary software

6.1 Introduction

The previous chapter focused on the ability of XAI to solve the problem of technical transparency. With this chapter, we start to move from a purely technical point of view to a sociotechnical and human-centered one. In particular, we examine the ability of XAI to audit a commercial black-box, i.e., an AI algorithm that is proprietary software (objective 3 of chapter 3). Indeed, the growing availability of Electronic Health Records (EHR) and the constantly increasing predictive power of Machine Learning (ML) models are boosting both research advances and the creation of business opportunities to deploy clinical Decision Support Systems (DSS) in healthcare facilities [181, 91, 242]. Since many of such models are not equipped to differentiate between correlation and causation, they might leverage spurious correlations and undesired biases to boost their performance. While there is an increasing interest in the AI community to commit to interdisciplinary endeavors to define, investigate and provide guidelines to tackle biases and fairness-related issues [275, 309, 259], quantitative and systematic auditing of real-world datasets and ML models is still in its infancy.

Ensuring the fairness of the suggestions provided by ML-based clinical DSS is a delicate task that requires to consider the whole process that goes from data to action. In critical scenarios, ML models do not make autonomous decisions without the supervision of a human; however, they might inadvertently learn to discriminate using unjustified bases for differentiation that reflect a history of systematically adverse outcomes for certain groups [29, 275, 281], thus leveraging and perpetuating harmful biases in their suggestions. Even under human supervision, the issue of biased suggestions of clinical DSSs is problematic since it has been shown that clinicians are affected by automation-bias, i.e., the tendency to over-rely on automation [134, 164, 221]. These findings highlight the importance of auditing the clinical DSS before it reaches its end-user.

While the source and the impact of errors of clinical DSS suggestions are numerous, in this chapter, we focus on errors that lead to systematic biases, and as consequence might cause fairness issues. In other words, we analyze the performance of a ML model across legally recognized protected groups such as gender, ethnicity, age, and on a proxy of socioeconomic status such as healthcare insurance. Indeed, model performance could create fairness issues if the algorithm suggestions on a protected group are systematically wrong [326, 259]. Our Research Question is therefore the following:

How can we audit a black-box Clinical Decision Support System in order to detect potential biases on different groups and explain its mislabellings on specific data points?

6.1.1 Bias in healthcare data and algorithms

Fairness issues can raise both from data biases and from biased algorithm [167, 344, 345]. However, in this chapter, we focus on fairness issues stemming from biased data. Indeed, healthcare data might contain several biases that can impact the model performance beyond its predictive accuracy. These biases are usually due to a lack of cohort diversity that might be originated by technical and non-technical reasons. Technical reasons that generate lack of cohort diversity:

- **Clinical study exclusion criteria** This happens when data used to train the DL model was collected for a specific target clinical population study, e.g., some studies focus only on adult population.

- **Poor data collection design** This generally applies every time the population used to train, validate and test the model does not reflect the target population of the clinical setting in which the model will be deployed [119, 342]. This mismatch might generate a wide range of biases. A comprehensive list of all these biases is outside the scope of this chapter, however, the main ones are *temporal biases* [60, 256, 161] (there is a concept drift between the time the model was developed and the time the model is deployed), *geographical biases* (the model was developed using only one-site data) [193], *bias due to confounding or omitted variables* (for example one missing variable such as the aggressivity of the treatment might mislead the model to wrongly classify high-risk patients as low-risk) [59] and *spectrum bias* (the population of the data set used to develop the model does not have a real representation of the spectrum of disease states – severity, stage etc. – of the target population) [273].

- **Secondary use of data collected with other purposes** An example of such practice is the use of ICD (International classification of Diseases) codes for predictive diagnosis purposes in DL applications. These codes were originally intended for billing purposes and might not properly describe the real health status of the patient [394, 261].

- **Lack of high-quality human labeling** This might happens for two reasons: the first one is the general low quality of the data set being used, the second one is specific to healthcare data. The problem arises from the fact that different doctors might give different diagnoses to the same patient. The fact that there is no gold standard for early cancer diagnosis well exemplifies this issue [312]. Furthermore, sometimes a lack of high-quality labeling might reflect a lack of knowledge: this is for example the case of sepsis prediction, there is no agreed upon definition of what sepsis is and thus, there is no universal ground truth [321].

Non-technical reasons of bias are due to the historical omission of certain populations from clinical studies [64, 326, 245, 6, 380, 206] and to the reflection of human biases and discrimination into the data set. Several examples in literature show how discriminatory biases influence ML outcomes. For example, Sayyed-Kalantari et al. [326] studied the bias of state-of-the-art Deep Convolutional Neural Network (CNN) on assigning the right diagnosis to chest X-ray images. They trained the CNN on three different large open data sets and showed that the underdiagnosis rate was consistently higher for women, minorities, and those with low socioeconomic status. Another work by Obermeyer et al. [259] exposes the racial bias of a risk-prediction algorithm used to rank patients according to their healthcare needs. They found out that using healthcare costs as a proxy label to identify patients that would benefit the most from targeted intervention was discriminating Black patients. This result was due to the fact that White patients generated higher healthcare costs conditional on health conditions with respect to Black patients, so the algorithm was favoring White patients. Even though most of the time fairness studies focus on legally protected groups, other forms of biases in healthcare could still be detrimental if ignored. For example, it is proven that many healthcare providers hold strong biases against people with obesity [289]. This attitude influences the quality of care provided and the healthcare outcomes of treatments [279]. Even if weight bias is still not regulated it could still be very harmful if silently perpetrated by ML applications in healthcare. Lastly, it is important to notice that removing sensitive features do not prevent discrimination since there might be other features correlated with the sensitive ones.

6.1.2 Fairness

The various stakeholders involved with the healthcare ecosystem (clinicians, patients/patient advocate, researchers, federal agencies and industry) identified the following urgent priorities for healthcare applications: trustworthiness, explainability, usability, transparency and fairness [90]. As suggested in [295], before launching (or deploying) a new ML-based product, a thoughtful auditing process is needed. While the auditing process involves multiple stakeholders and embrace several as-

pects of product development, one of the ultimate goals is to help understanding if the ML model outcomes are fair. Consequently, the auditing process helps to choose the best actions to perform or the best bias mitigation strategy to adopt. Building an auditing system first requires defining fairness according to societal values and then operationalize it. Many efforts have been devoted to detecting and measuring discrimination in model decisions [308, 398, 150]. Several definitions and methodologies have been proposed to measure bias and fairness [275, 100, 229, 152]; however, despite the effort, a general consensus on such measures is still missing. This is because the most appropriate fairness metric is highly context-dependent. Generally speaking, the most prevalent approach to fairness in machine learning is to solicit for approximate parity of some statistics of the predictions (such as false negative rate) across pre-defined groups [200, 194, 76]. Moreover, there are very few available general-purposes resources to operationalize them [12, 354, 35, 309]. The majority of such research has focused on binary or multi-class classification problems to prevent discrimination based on sensitive attributes assessing fairness issues between only two groups (e.g. female vs male, black vs white) [116], and a few studies focus specifically on multi-label classification problems, which is the learning problem of the presented FairLens use-case, with many concentrating on fairness in ranking and recommendation systems [8, 101, 126]. In the context of medical applications, a recent paper [65] suggested that the post-deployment inspection of model performance on groups and outcomes should be one out of five ethical pillars for equitable ML in the advancement of health care.

6.2 Main contribution

This chapter is a based on our paper:

- Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657, 2021

In this paper we introduced **FairLens**, a methodology for discovering and explaining biases. We show how this tool can audit a fictional commercial black-box

model acting as a clinical DSS (DSS). In this scenario, the healthcare facility experts can use FairLens on their historical data to discover the biases of the model before incorporating it into the clinical decision flow. FairLens first stratifies the available patient data according to demographic attributes such as age, ethnicity, gender and healthcare insurance; it then assesses the model performance on such groups highlighting the most common misclassifications. Finally, FairLens allows the expert to examine one misclassification of interest by explaining which elements of the affected patients' clinical history drive the model error in the problematic group. We validate FairLens' ability to highlight bias in multilabel clinical DSSs introducing a multilabel-appropriate metric of disparity and proving its efficacy against other standard metrics.

6.3 FairLens: target user and context

FairLens is an auditing tool that allows to test a clinical DSS before its deployment, i.e., before handing it to final decision-makers such as physicians and nurses. The designated user of FairLens is a healthcare facility expert who wants to audit the ML model before adopting and deploying it in the facility, as illustrated in Figure 6-1.

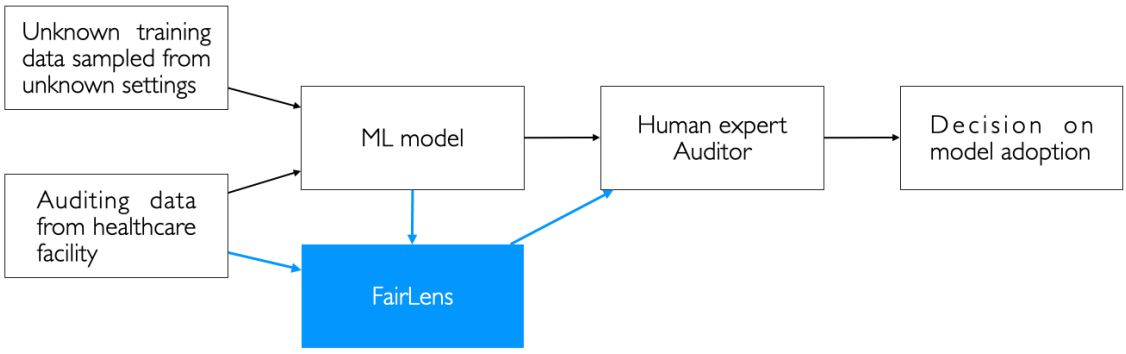


Figure 6-1: FairLens as a tool for auditing a clinical decision support system before its deployment in a healthcare facility. Our contribution, in blue, provides to the auditor an instrument to detect and explain systematic ML model biases on protected groups.

In this representation, the origin of model bias might be either in the unknown training data or in the learning process. Since it is generally not possible to access the training data used to build the clinical DSS, FairLens can become a powerful tool

to assess if the model is appropriate for the specific hospital's reference population, i.e., the auditing data of Figure 6-1. Indeed, FairLens allows the human expert to perform a thorough analysis of potential fairness issues. However, the final decision on whether the signaled bias constitutes a real problem or it is a justified basis for differentiation is left to the auditor. Ideally, the FairLens user is an IT expert with a quantitative background and an in-depth knowledge of the healthcare setting, for example, the director of the IT department of a big hospital. This type of user usually has the responsibility to ensure the quality and trustworthiness of new technologies before adoption. FairLens then becomes an additional tool to understand whether to adopt the system or to evaluate if a bias mitigation strategy is needed, for example, by post-processing the DSS outcomes.

FairLens takes bias analysis a step further by explaining the reasons behind the poor model performance on specific groups. FairLens embeds explainability techniques in order to explain the reasons behind model mistakes instead of simply reporting model scores.

Throughout this chapter, we present a use case where FairLens is used to investigate the potential biases in ML models trained on patients' clinical history represented as diagnostic codes using the *International Classification of Diseases* (ICD) standard. This type of structured data allows for a machine-readable representation of the patient's clinical history and is commonly used in longitudinal ML modeling for phenotyping, multi-morbidity diagnosis classification and sequential clinical events prediction [390, 72, 63]. As already mentioned in section 5.1.5, the implicit assumption behind the use of ICD codes in this kind of ML applications is that these codes are a good proxy for the patient's actual health status. However, ICD codes can misrepresent such status due to many potential sources of error in translating the patient's actual disease into the respective codes [261, 66]. This is particularly true when ICD codes are fed into *black-box* ML models, i.e., models whose internal decision-making process is opaque.

6.4 FairLens: pipeline

This section describes the FairLens methodology to audit black-box clinical DSSs in order to *i*) detect potential biases on different groups and *ii*) explain its mislabellings on specific data points. Here we describe an end-to-end use of FairLens on a specific setting (i.e. prediction of future health conditions, based on past observation of ICD codes), and we provide an alternative scenario in the Appendix A. Indeed, it is worth stressing that the functional blocks of the pipeline are quite general and thus FairLens can also be used in different settings after an appropriate tailoring of the modules. In particular, different applications might be interested in stratifying the data points according to different categories other than gender, ethnicity, age and insurance. Moreover, according to the classification problem at hand, the scoring measure might be different from the one presented here for the high-dimensional multi-label classification, and clearly the explanation method should be suitable for the black-box as well. Such considerations highlight the potential of FairLens as a useful framework to allow humans inspecting algorithmic decision-making pipelines, without delegating to yet another automated tool the delicate task of auditing unintended and potentially harmful consequences of decision support systems. As such, our approach provides insights about the *who* and the *why* of the differential treatment of a clinical DSS on certain groups, letting the human experts understanding if such behaviour is legit or may lead to fairness issues.

Given a black-box to audit, the building boxes of the pipeline described hereafter are: stratification, scoring, ranking, inspection, explanation and summary report. A bird's-eye view of the pipeline is depicted in Figure 6-2.

Let BB be a sequential black-box ML model trained on ICD data. The model can be available as an on-premise-installed software or it could be integrated via an exposed API. The only requirement about BB is that it can be queried at will. Let $P = \{p_1, \cdots p_N\}$ be the set of patients. Let each patient p_i be represented as (p_i^{att}, p_i^{ch}), where p_i^{att} is a set of attributes such as *ethnicity*, *gender*, and *insurance type*, and $p_i^{ch} = \{v_{i,1}, \cdots v_{i,V}\}$ is the clinical history represented as a sequence of visits. In turn, each visit is represented by a set of ICD codes. Let

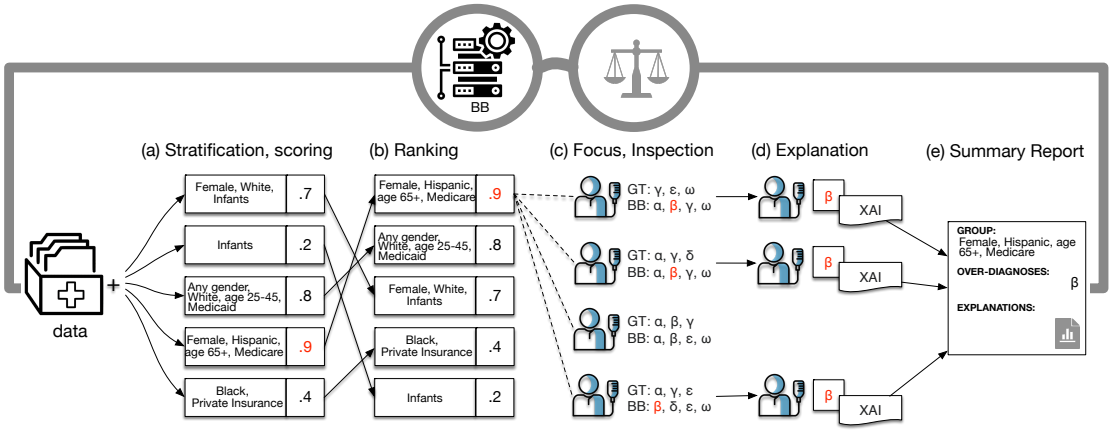


Figure 6-2: Fairlens pipeline: a tool to support human experts investigating if a black-box clinical DSS behaves differently on groups based on protected attributes, highlighting which health conditions are more often misclassified and why.

$v_{i,j}^{BB} = BB(\{v_{i,1}, \cdots v_{i,j-1}\})$ be the prediction of the black-box for the $j-th$ visit of patient p_i.

It is worth to notice that the p_i^{att} are not part of the feature space of the BB, and in principle the patients' attributes could be more than those presented here to exemplify the use of FairLens. In general, p_i^{att} could include any attribute that is not used by the model to predict future health conditions, but can be collected in a structured database (e.g. education level, job status).

Stratification

The first step of our methodology is depicted in Figure 6-2(a). Since we aim to compare the ML model performance across groups, we stratify our patients set P according to a set of conditions c on the set of attributes p^{att}, e.g. $c = \{\text{age} \leq 40, \text{insurance} = \text{Medicaid}\}$. Given a set c of conditions, we define a *group* G as the set of non-first visits of each patients whose attributes match the conditions in c:

$$G_k = \{v_{i,j} \mid j > 1, v_{i,j} \in p_i^{ch}, p_i \in P, p_i^{att} \in c_k\}$$

The stratification process produces a set of groups $G_1, .., G_M$. While the stratification process is based on the attributes of patients, we create different data-points for each non-first visit, so that we can evaluate the performance of the model on

every visit of the patients' clinical history. Considering each visit as a different data point is necessary because some demographic attributes might change between two visits of the same patient (consider for example *age* and *healthcare insurance*). The first visit of each patient ($j = 1$) are excluded because in those cases the model has no previous patient history to base its prediction upon.

We remark that there is a degree of freedom regarding which set of attributes are considered. The granularity might be tuned at will, ranging from one-attribute constraints $\{\text{gender} = \text{F}\}$ to more detailed constraints $\{\text{gender} = \text{F}, \text{age} \geq 65, \text{ethnicity} = \text{white}, \text{insurance} = \text{Medicare}\}$. A domain expert might suggest specific condition sets to isolate a given sub-cohort of known interest, whereas a technician might opt for building a lattice of all possible combinations of constraints. The attributes considered here are deemed relevant for auditing purposes as existing literature suggests that minority groups might be at risk of fairness issues, and protected attributes (i.e. traits or characteristics that, by law, cannot be discriminated against as age and gender) should not affect the model performance. Here, we also considered the insurance type as it is a proxy for socioeconomic status. According to data availability, other attributes could be further added to the stratification process. We also remark that some patients might not occur in any group or occur in more than one, depending on the provided conditions.

Scoring

After the stratification step, FairLens proceeds to the scoring phase. For every non-first visit $v_{i,j}$ occurring in any group G_k, we query the BB on the previous clinical history of that patient, so that we can compare the ground-truth visit $v_{i,j}$ with its predicted counterpart $v_{i,j}^{BB} = BB(\{v_{i,1}, \cdots v_{i,j-1}\})$. We therefore obtain the predicted counterparts for every visit in every group, and we can evaluate how different groups fare in terms of truth-prediction disparity.

Although many works in literature define disparity as a distance according to a reference group [195], here we choose to define disparity as a measure relative to a target standard, that in the case of ML algorithms might be e.g. perfect prediction of the target values. Therefore, for the purposes of this discussion, we propose the

following definition of disparity:

The quantity that separates a group from a target standard using a particular measure of performance.

Hence, a *disparity function* $d : G_k \rightarrow s_k$ maps every group G_k to a *disparity score* s_k. FairLens includes a number of disparity functions, such as the standard classification metrics (such as accuracy and F1-score) and distribution-comparison functions like the Wasserstein distance. Custom disparity functions can be used, as long as their results can be used for ranking. Given a disparity function, FairLens computes the score s_k for each group G_k, which represents the performance of the BB on that specific set of patients.

Ranking

Once each group has been scored, FairLens ranks the groups, as depicted in Figure 6-2(b). The ranking highlights groups where the BB performs relatively poorly, signaling them to domain experts for further inspection. Alternatively, the domain experts might arbitrarily select one group for further inspection, regardless of their scores, due to the cohort's known peculiarities or clinical-dependent reasons.

Inspection

Given a specific group G_k flagged for further inspection by the group ranking function, FairLens compares the black-box prediction $v_{i,j}^{BB}$ with the ground truth $v_{i,j}$ for each visit in G_k. The goal of this step is to check for systematic bias of the BB on the group of patients. For each diagnostic code, the relative frequencies in the predicted and true values are computed and we define the *misdiagnosis score* the difference between these two values. Ranking the codes by misdiagnosis scores allows to highlight which diagnostic codes are particularly over- or under-predicted (high and low difference values respectively). FairLens thus displays the top three over- and under-represented codes to the domain expert who can ask for an explanation for the highlighted conditions that might result in producing or reinforcing systematic over- or under-treatment. In Figure 6-2(c), we have labelled the true

visit value as *GT* (for *ground truth*); in the mock example it can be observed that the code β is over-represented.

Explanation

In order to extract an explanation for the mislabeled code, FairLens first assigns binary labels on the visits of the group of interest. Suppose the domain expert wants to understand what elements of the group clinical histories are most influencing the over-representation of ICD code β in the inspected group G_k, then at each visit $v_{i,j} \in G_k$ will be assigned a binary label representing the misclassification of the ICD code β:

$$l(v_{i,j}) = \begin{cases} 1 & \text{if } (\beta \in v_{i,j}^{BB}) \oplus (\beta \in v_{i,j}) \\ 0 & \text{if } (\beta \in v_{i,j}^{BB}) == (\beta \in v_{i,j}) \end{cases}$$

Then, FairLens selects all the misclassified visits (binary label 1) and explains them using a local XAI technique for sequential healthcare data. Typically XAI techniques are used to explain the outcome of a black-box ML model. In this setting, we want to explain why the specific code was wrongly assigned, and we do so by providing the XAI technique with the custom binary label.

More generally, we define the *Explainer* as a function:

$$\xi : (BB, x_i, \beta) \rightarrow \{\ f_1 \gtrless t_1 \cdots, f_F \gtrless t_F\}$$

that maps a blackbox BB, a patient's feature vector x_i and a clinical code β to a set of decision rule premises $\{\ f_1 \gtrless t_1 \cdots, f_F \gtrless t_F\}$ where each f is a feature in x_i that, in combination with a threshold value t, explains why BB misclassified β for the patient p_i. In the case where a black-box BB predicts β from a feature vector x_i that is the patient's clinical history p_i^{ch}, the feature names f are a subset of the medical codes in p_i^{ch}.

It is worth noting that while XAI techniques are usually employed to explain the reasons behind a black-box decision, thanks to the aforementioned binarization process, FairLens uses them to explain the reasons behind a specific mislabelling.

Furthermore, we observe that when a model-agnostic XAI technique is employed, FairLens can be used audit any model without having access to its internal structure or parameters. However, FairLens can be used with model-aware XAI techniques too, and we provide an example in the Supplementary Information.

Reporting

Finally, FairLens combines the local explanations of each mislabelled visit of group G_k in one set of *global* rules; this corresponds to step (e) in Figure 6-2. The local explanations extracted by FairLens are in the form of decision rules with premises. Each condition of the rule premise follows the pattern

$$\text{ICD\_code} \gtrless threshold\_value$$

where the *threshold value* expresses whether and when the ICD code was observed in the patient's clinical history. These local explanations are merged by FairLens employing a state-of-the-art XAI technique, GlocalX [325], that outputs a compact set of global rules by hierarchically merging the local explanations based on their similarity. Finally, FairLens translates the final set of global rules into natural language and presents the report to the user.

6.5 Use Case: auditing a medical decision support system

In this section we show how a domain expert can use FairLens on the historical data available at her healthcare facility to audit a fictional commercial clinical *decision support system* (DSS) that predicts patient's future clinical events based on their clinical history. We assume that the domain expert has access to the DSS as a *black-box*, i.e. she can query the DSS at will but has no access to its source code, to its weights or to the data used for its training. We use the MIMIC-IV (see Subsection 6.5.1) database of electronic health records as the fictitious historical database of the facility and DoctorAI (see Subsection 6.5.2) as the fictional clinical DSS. We

split the dataset in training (29.714 patients, 68%), validation (5.244 patients, 12%) and test set (8.739, 20%). Training and validation sets are used to deploy DoctorAI as a black-box and are not seen during the auditing process, while the patients in the test set are used as auditing data. We exploit DoctorXAI (Subsection 6.5.3) as the backbone of the FairLens explainer, and we show how this auditing process is effective to detect and explain potential biases on different groups.

6.5.1 Dataset: MIMIC-IV

The MIMIC (Medical Information Mart for Intensive Care) [135, 188] database is a single-center freely available database containing de-identified clinical data of patients admitted to the ICU (intensive care unit) of the Beth Israel Deaconess Medical Center in Boston. Its most recent update, MIMIC-IV [187], contains information of 383,220 patients collected between 2008 and 2019 for a total of 524,520 hospital admissions. The database includes patient's demographics, clinical measurements and diagnoses and procedures codes of each admission. We focused our analysis on hospital admissions coded with ICD-9 billing codes and on patients having at least two admissions to the hospital, reducing the number of patients to 43,697 and the number of admission to the hospital to 164,411 (see table 6.1).

number of patients	43,697
number of admissions	164,411
avg. nr. of admissions per patient	3.76
max nr. of admissions per patient	146
number of unique ICD-9 codes	8,259
avg. nr. of codes per admission	11.22

Table 6.1: MIMIC-IV: Data from patients with at least two hospital admissions

6.5.2 Clinical DSS: Doctor AI

Doctor AI by [72] is a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU) that predicts the patient's next clinical event's time, diagnoses and medications. For the purpose of this use-case, we focused only on diagnoses prediction. We trained the model on MIMIC-IV using the training and validation set

as defined previously using default hyperparameters. Doctor AI can be trained to predict patient's future clinical event in terms of either CCS (Clinical Classifications Software) or ICD codes. CCS codes are used to group ICD codes into smaller number of clinically meaningful categories. As suggested in [72] we trained Doctor AI to estimate the probability that a CCS code is assigned to a visit at time $t+1$ given the ICD-9 codes assigned to patient's visits until time t, and measured its performance using Recall@n with $n = 10, 20, 30$.

6.5.3 Local Explainer: DoctorXAI

DoctorXAI [271] is a post-hoc explainer that can deal with any multi-label sequential model. Since it is agnostic w.r.t. the model, i.e. it does not use any of its internal parameter in the explanation process, it is suitable for our methodology which considers the clinical DSS as a black-box. Furthermore, DoctorXAI exploits medical ontologies in the explanation process and in our case we exploited the ICD-9 ontology. The explanations provided by DoctorXAI are *local* decision rules, which means that they provide the rationale for one particular classification. In our scenario, we want to provide an explanation for a over- or under-diagnosis observed in a group of patients, therefore FairLens binarizes the black-box probability estimates and it combines the explanations as described in the *Explanation* and *Reporting* paragraphs of Section 6.4.

6.5.4 Local-to-global approach: GlocalX

GlocalX [325] is a model-agnostic XAI algorithm that explains the global behavior of black-boxes by aggregating a set of local explanations in the form of decision rules. GlocalX hierarchically merges local explanations optimizing both the complexity and fidelity of the decision rules set, i.e., its size and ability to mimic the black-box behavior correctly. In our case, we used GlocalX to merge all the local explanations extracted by DoctorXAI to explain the individual misclassifications of a group. We stress that while GlocalX is a methodology to generate a transparent model able to mimic the black box's global behavior, in our scenario, we use it as an aggregator of

explanations for the patients of the group under investigation, i.e., all the patients having a specific misclassification. Therefore the validity of the provided global explanation is limited to the black-box behavior on those patients.

As described in Section 6.4, DoctorXAI produces rules that follow the pattern ICD_code $\gtrless$ *threshold_value*, and GlocalX preserves this structure. To map back these rules onto human-readable sentences, we simply revert DoctorXAI's temporal encoding. In order to circumvent the temporal nature of medical history data, DoctorXAI exploits a fairly straightforward temporal encoding, where each ICD9 code receives an exponentially decreasing value according to its occurrence (or lack thereof) in the visits of the patient, explored backwards. The last visit corresponds to a score of .5, the second-to-last to a score of .25, and so on. For instance, if some condition C was diagnosed in the third-to-last and second-to-last visits, but not in the last one, C would be given the value of .375. Given this logic, it is trivial to interpret the inequalities produced by DoctorXAI and aggregated by GlocalX: $C < .5$ means, for instance, that the ICD9 C was not diagnosed in the last visit, while $C >= .25$ means that the ICD9 C was diagnosed at least once in the last two visits of the patient.

6.5.5 Auditing DoctorAI on MIMIC-IV

Assessing the DSS performance on the healthcare facility data. The first step that a domain expert would perform before deploying the clinical DSS on her dataset is to measure its global performance on the facility data. In our scenario, a domain expert would obtain the results in table A.1.

BB Recall	**@10**	**@20**	**@30**
On auditing data	0.481	0.623	0.712

Table 6.2: clinical DSS performance

Identify problematic groups of patients

Once the global performance has been assessed, the domain expert can apply FairLens to discover potential biases learned by the model. The domain expert would

start by deciding which attributes to use to stratify the patients. For the purpose of our fictional scenario, we consider the following attributes occurring in the auditing data: *Gender*, *Ethnicity*, *Age* and *Insurance type*. The distributions of these attributes is shown in figure 6-3.

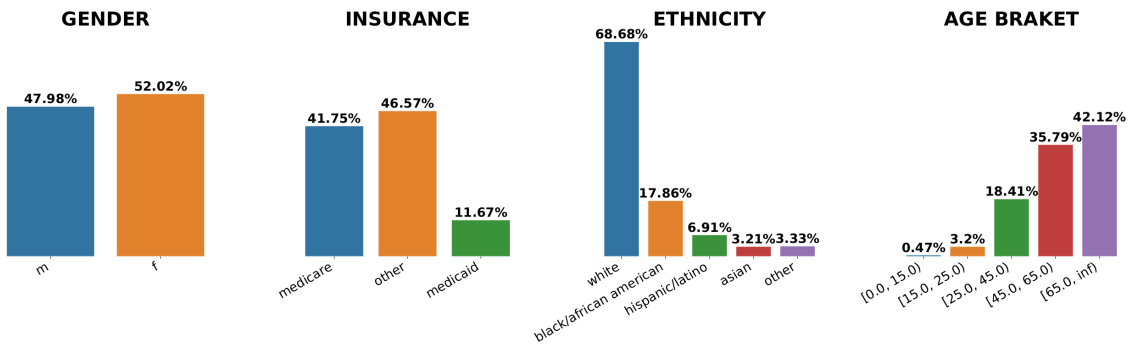


Figure 6-3: Distributions of demographic attributes in the auditing data

Once these attributes are selected, FairLens computes the disparities across groups. In our scenario, the black-box is a sequential multi-label model that predicts the set of codes diagnosed in the next visit in terms of CCS codes. In this multi-label case, the disparity is evaluated using the *Wasserstein distance* which has already been successfully employed as a loss function for multi-label and multi-class ML tasks [123] and to post-process the output of a classifier to achieve fair treatment [182]. This metric measures the distance between two probability distributions: for each group of interest, the distance between the distribution of CCS codes in the black-box output and the same distribution in the ground truth. In our scenario, the DSS outputs the top 30 CCS codes ranked by estimated probability. Similarly to the *recall@k* we define the *disparity score@k* which is the Wasserstein distance between the ground truth and the predicted probability distributions over the top-k CCS codes. From now on, we will perform the analysis using the *disparity@30* unless otherwise specified.

The domain expert can decide to either explore a specific group of interest or to have a comprehensive view of the biases of the DSS on all possible groups.

The scatter plots in Figure 6-4 confront the normalized disparity score with the group size for all possible groups. Each scatter plot focus on a specific attribute,

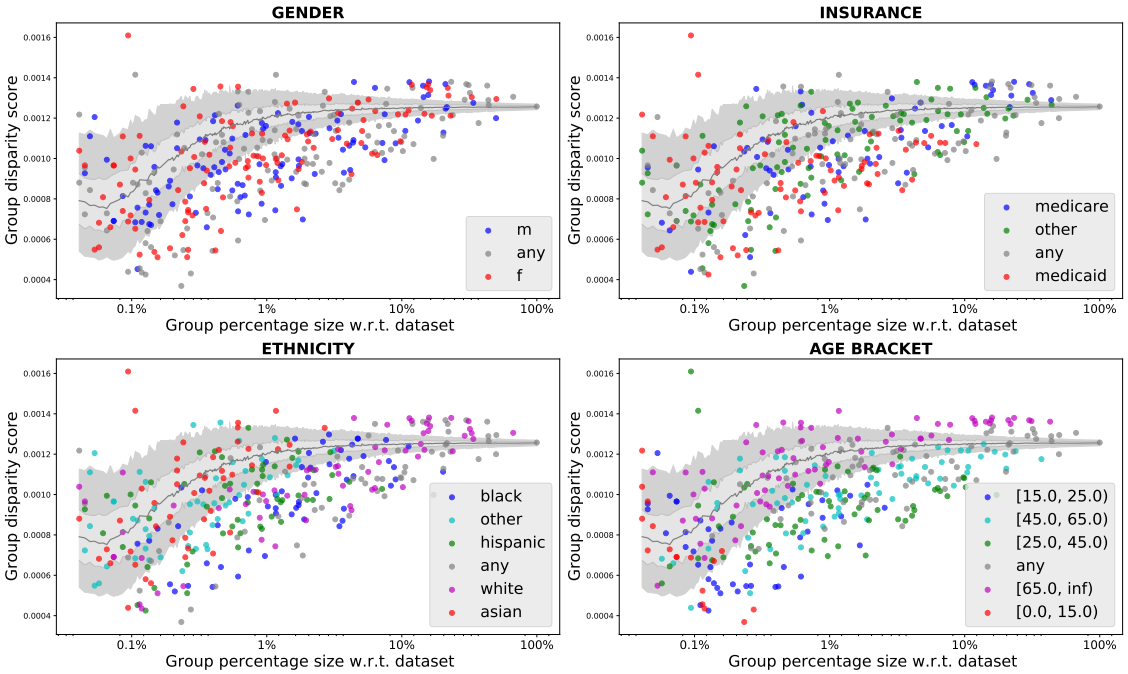


Figure 6-4: Normalized disparity scores vs. group sizes with bootstrap outliers bands capturing 50% (light grey) and 95% (dark grey) of the random variability for that group size. The median of the bootstrap distribution is shown as a solid grey line.

and each point represents a group with a combination of attributes, for a total of 340 combinations. The color-coding allows to explore the disparities of each intersectional identity. Data points labelled and color-coded as *any* correspond to groups that do not represent a specific value for the stratification feature: for instance, the group *(male, medicare)* includes patients of all ages and ethnicities.

In the same plots we show the variability in disparity score as function of the group size, when selecting the same number of patients independently of their group assignment. In particular, we randomly sample with replacement 1000 times for each group size to estimate its disparity score's sampling distribution. The plots show the median (solid grey line) and the bands capturing the 50% (light grey) and 95% (dark grey) of the distribution for each group size. The groups falling above the dark grey band's upper limit have a disparity score above the 95th percentile of the distribution for that group size when no demographic variable is considered. These groups are also marked with an asterisk in Table 6.3.

A higher variability in terms of disparities is observed among smaller groups.

Group size bin	Insurance	Gender	Age group	Ethnicity	Disparity score	Group size
10-50	medicaid	f	25-45	asian	1.00 ∗	23
50-100	medicare	f	over 65	other	0.84 ∗	70
100-200	any	f	over 65	other	0.84 ∗	111
200-400	any	any	over 65	asian	0.88 ∗	286
400-800	any	any	any	asian	0.83 ∗	657
800-1500	other	m	over 65	white	0.86 ∗	1082
1500-3000	medicare	m	over 65	white	0.86 ∗	2783
3000-5000	any	m	over 65	white	0.86 ∗	3894
5000-24446	medicare	any	over 65	white	0.86 ∗	5679

Table 6.3: Groups with the highest disparity score in each group size bin. All disparity scores marked with ∗ are above the 95th percentile of random variability for the group size.

While this might suggest fairness issues for relatively rare groups, given the small size of these groups, the high variability and dispersion away from the mean could also occur by chance; therefore Table 6.3 also provides an overview of the groups with the highest disparity in predefined group-size bins. The results in this table (supported by the Age Bracked plot in Figure 6-4) suggest that the DSS seems to often misdiagnose older patients; indeed they are the most prevalent age group with the largest disparity score by group size bin.

Identifying systematic sources of error in the selected group.

For each group, FairLens then computes the misdiagnosis score of each CCS code by subtracting its ground truth value (clinical conditions) from the value predicted by the DSS. This score allows to rank the codes, so that the the most over- and under-diagnosed CCS codes can be isolated. Table A.2 reports the top 3 groups by disparity score in the largest bins, and the top 3 codes ranked by over- and under-diagnosis scores.

The domain expert auditing the system can further select a specific group for a more in-depth investigation. Suppose she decides to focus on one of the groups with the highest disparity and also a fairly high group-size, for example patients of Asian ethnicity and over 65 years of age (see Table A.2). This analysis tells the domain expert that across groups the DSS tends to over-diagnose general conditions such as *Essential hypertension* or *Unclassified*. More interestingly, for the group of patients of Asian ethnicity and over 65 years of age, the DSS seems to under-diagnose *Heart*

Group	Size	Disp. Score	Over-diagnosed	(Misdiagnosis Score)	Under-diagnosed	(Misdiagnosis Score)
Female, 65+, Medicare, Other ethn.	70	0.83	106: Dysrhythmia	0.027	2621: E Codes: Place of occurrence	-0.010
			98: Essential hypertension	0.02	2603: E Codes: Fall	-0.009
			259: Unclassified	0.019	210: Systemic lupus erythematosus	-0.007
Female, 65+, Other ethn.	111	0.84	259: Unclassified	0.024	2621: E Codes: Place of occurrence	-0.009
			98: Essential hypertension	0.024	2603: E Codes: Fall	-0.007
			106: Dysrhythmia	0.022	250: Nausea/vomit	-0.006
Asian, 65+	286	0.88	259: Unclassified	0.023	6: Hepatitis	-0.009
			98: Essential hypertension	0.020	204: Other non-traumatic joint disorder	-0.008
			663: Hist. of mental health and subs. abuse	0.016	96: Heart valve disorders	-0.007

Table 6.4: Groups ranked by disparity scores and most over/under-diagnosed conditions when auditing the black-box

valve disorder, which is a potentially severe condition that might need surgery.

Obtaining explanations for systematic misclassifications.

Once the groups with the highest disparities are identified, the domain expert can use FairLens to obtain an explanation for one particular misclassification. Consider, for example, the under-diagnosis of *Heart valve disorders* (CCS code 96) for over-65 Asian patients. FairLens uses DoctorXAI to discover which elements in the patients' clinical history drive the under-diagnosis of that specific CCS. This is done by first projecting the black box's multi-label output on the single label 96 (as explained in Section 6.4), then calling DoctorXAI to explain the binarized outcome for the 19 patients where the CCS code 96 was wrongly not diagnosed. By doing so we obtain 19 explanations, one for each CCS-96-misdiagnosed patient in our patients group. As a further step, the GlocalX local-to-global algorithm aggregates these local explanations into a more compact and doctor-readable global explanation. GlocalX, for this explanation set, produces the global rules of Table 6.5.

While the original rule set had 19 rules of mean length 10, the resulting rule set contains only 5 rules of mean length 8. Clearly, this is a more compact set but not yet comprehensible.

As a very first feedback to the expert, FairLens produces Figure 6-5: this plot highlights the ICD9 codes that occur in the global rules (and therefore are brought out by the FairLens pipeline as misclassification culprits) and are also most common among the patients of the group under scrutiny. In our case, for instance, the domain

427.31<=0.25	**410.91**<=0.25	**396.3**<=0.25	**410.71**<=0.25	**424.0**>0.25	**162.3**>0.125
424.1<=0.25 **V10.3** > 0.004	**425.4** > 0.16 **V49.86** > 0.024	**202.10** > 0.0005 **V12.72** > 0.033	**427.31** > 0.5 **E930.7** > 0.016	**244.9**>0.5 **V45.82**>0.244	**E933.1**>0.1
427.31 > 0.62 **362.50** > 0.125 **V04.81** > 0.25	**V45.82** > 0.125 **530.81** > 0.375	**428.0** > 0.437 **411.1** > 0.25	**567.29** > 0.125 **412** > 0.187	**575.4**>0.125 **401.9**>0.25	**574.00**>0.062 **564.00**>0.062
427.31<=0.25					
424.1 > 0.25 **288.3** > 0.0625	**V12.71** > 0.344 **255.9** > 0.0625	**401.9** > 0.148 **V13.01** > 0.25	**305.1** > 0.219	**E849.9**>0.023	**403.90**>0.344

Table 6.5: Set of rules produced by DoctorXAI and aggregated by GlocalX to explain why the CCS code 96: *Heart valve disorders* was under diagnosed for over-65 Asian patients by the model DoctorAI. Each row group is a rule with a set of premises, each premise is in the form of ICD-9 $\gtrless$ *threshold_value*. For the human-readable description of each ICD-9 code the reader can consult http://www.icd9data.com/.

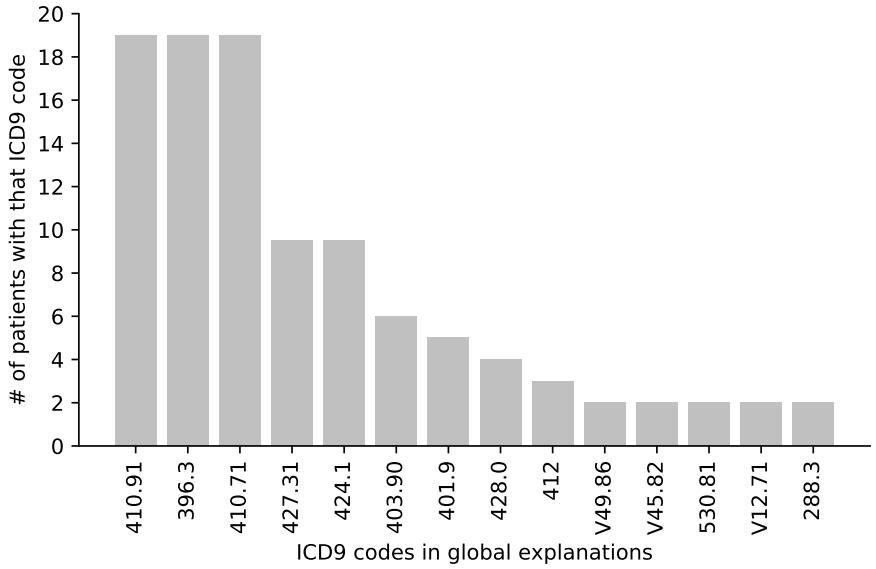


Figure 6-5: Aggregated visualization of the relevant ICD-9 codes for the under-diagnosis of *Heart valve disorders* in over-65 Asian patients.

expert can immediately observe that the highlighted ICD9 codes are 410.91 (*Acute myocardial infarction of unspecified site, initial episode of care*), 396.3 (*Mitral valve insufficiency and aortic valve insufficiency*) and 410.71 (*Subendocardial infarction, initial episode of care*).

Figure 6-5 provides useful preliminary insights to the FairLens user, but at the same time the information conveyed by the global explanations is richer and can be presented in greater details. First, we want to translate these rules back into natural language, and we do so as explained in the previous subsection: for instance, the

last global conjunct is **V13.01**>0.25 and it corresponds to *'Personal history of urinary calculi' was diagnosed in the last visit.* Second, we want to rank our global rules. To do so, we measure the coverage of each rule as the number of patients whose features do not violate the rule, and we select the rules in a greedy fashion, highlighting those with higher coverage. For our case-study, the re-interpreted output of GlocalX is the following:

- FairLens focused on 19 patients
- 13 patients were misdiagnosed because *'Atrial fibrillation'* was not diagnosed in the last visit.
- 5 remaining patients were misdiagnosed because *'Aortic valve disorders'* was not diagnosed in the last visit, *'Other primary cardiomyopathies'* was diagnosed at least once in the latest two visits, *'Mycosis fungoides, unspecified site, extranodal and solid organ sites'* was diagnosed at least once in the latest three visits, *'Atrial fibrillation'* was diagnosed in the last visit, *'Unspecified acquired hypothyroidism'* was diagnosed in the last visit, *'Antineoplastic and immunosuppressive drugs causing adverse effects in therapeutic use'* was diagnosed at least once in the latest three visits, *'Personal history of malignant neoplasm of breast'* was diagnosed at least once in the latest three visits, *'Do not resuscitate status'* was diagnosed at least once in the latest three visits, *'Personal history of colonic polyps'* was diagnosed at least once in the latest three visits, *'Antineoplastic antibiotics causing adverse effects in therapeutic use'* was diagnosed at least once in the latest three visits, and *'Percutaneous transluminal coronary angioplasty status'* was diagnosed at least once in the latest two visits.
- 1 remaining patient was misdiagnosed because *'Aortic valve disorders'* was diagnosed in the last visit, *'Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified'* was diagnosed in the last visit, and *'Eosinophilia'* was diagnosed at least once in the latest three visits.

This human-readable snippet is the final output of FairLens pipeline - it provides medical experts with insights on why the medical decision support system misdiagnosed patients of the selected group, failing to diagnose the highlighted condition, CCS 96 - *Heart valve disorder.*

6.6 Validation

To empirically validate the reliability of FairLens in discovering biases, we created an artificially biased DSS and we ran the FairLens pipeline on it. The aim of this validation is to check whether the disparity measure used by FairLens is able to highlight the bias we injected in the DSS even when standard measures of multi-label performance (e.g. *recall@n* and *microAUC)* do not detect it.

Creating the biased DSS.

One of the most common causes of bias in machine learning is the under-representation of some categories in the training set. We then performed a random undersampling of patients having *Other* as Insurance, removing 90% of them from MIMIC-IV dataset (sampling A of figure 6-6). Finally, we used this skewed dataset as the training set for DoctorAI creating the biased DSS. While in this case we used such approach to validate the proposed pipeline, it is worth to notice that several studies suggest that ICD9 codes might be severely biased by the insurance type variable [282, 154, 128, 230].

The biased DSS created using this training set also contains, by construction, all the biases already present in the original dataset. To check whether the bias detected by FairLens in the biased DSS is actually the one we synthetically injected rather than the one already present in the original dataset, we created a *baseline* DSS by training DoctorAI on a random undersampling of MIMIC-IV (sampling B in figure 6-6). This sampling creates a training set that has the same size as the biased one, but that has the same distributions of demographic variables as the auditing dataset. Figure 6-6 shows the resulting distributions of training set demographic variables for the two sampling and for the test set.

The fact that the size of the training set is the same for both the biased and the baseline DSS allows a fair comparison of the performance metrics among the two. Indeed comparing the performance of the biased DSS with a baseline trained on a MIMIC-IV dataset withoug sampling would result in a baseline performance higher than the biased one only due to the bigger size of training set, creating a confounding factor for the analysis. The performance of the two DSSs on the test set are shown in Table 6.6.

Comparing the distributions of demographic variables of these two black-boxes (Figure 6-6), we note that by removing 90% of patients having *Other* insurance, we also changed

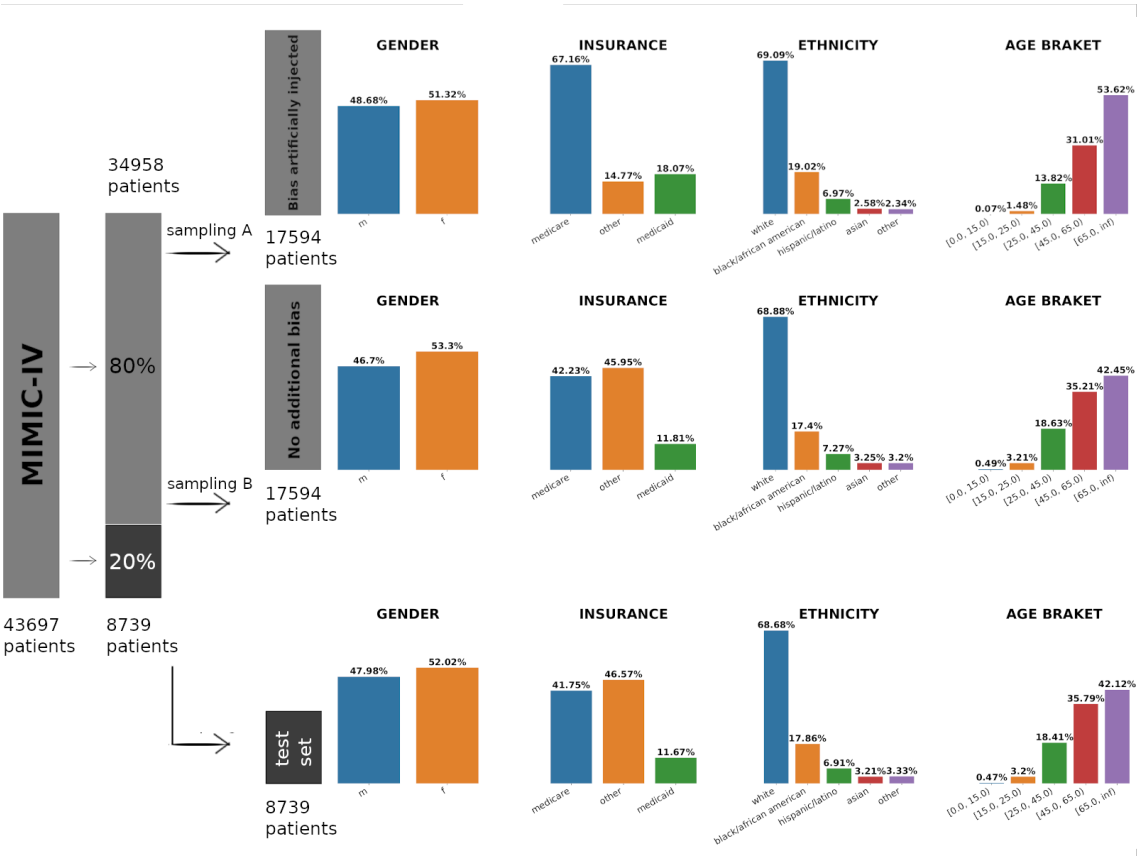


Figure 6-6: Sampling procedure and distribution of demographic variables in the training sets and test set. We first extract from the whole MIMIC-IV a test set with 20% of data points. From the remaining points, we extract two training sets with different sampling procedures (sampling A and sampling B). Sampling A produces a training set with artificially injected bias. Sampling B produces a training set with random sampling that respects the same distribution of demographic variables as the original dataset.

BB Recall trained	**@10**	**@20**	**@30**
on biased training set	0.449	0.586	0.671
on baseline training set	0.454	0.591	0.683

Table 6.6: Performance of clinical DSS trained on the biased and on the baseline training sets

the distributions of other demographic variables. Consider, for example, the age distribution in the biased training set. We can see that patients having age $0-15$ almost disappear from the dataset.

6.6.1 FairLens Analysis

We then proceed to run FairLens Pipeline on these two DSS. The first step is to identify potentially problematic groups of patients using FairLens scatterplots (Figure 6-7) and tables (Table 6.7).

Comparing the two scatterplots we can immediately see that FairLens detect both the *Insurance* and the *Age* bias synthetically injected in the biased DSS. Indeed, the majority of patients having the biggest disparity scores are those of age 0-15 and those having insurance *Other*. This is visible also in the tables that show the highest disparity scores binned by group size (see Tables 6.3 and 6.7).

group size bin	insurance	gender	age braket	ethnicity	disparity score	group size
10-50	other	f	0.0-15.0	white	1.00 ∗	10
50-100	other	any	0.0-15.0	any	0.66 ∗	57
100-200	other	m	any	asian	0.42 ∗	149
200-400	any	any	over 65	asian	0.40 ∗	286
400-800	any	any	any	asian	0.39	657
800-1500	other	m	any	black/african american	0.40 ∗	866
1500-3000	medicare	f	over 65	white	0.41 ∗	2896
3000-5000	medicare	f	over 65	any	0.41 ∗	3832
5000-24447	any	f	over 65	any	0.40 ∗	5351

Table 6.7: Groups with the highest disparity score in each group size bin for the biased DSS. All disparity scores marked with ∗ are above the 95th percentile of random variability for the group size.

insurance	**rank on baseline**	**rank on biased**	**mean rank on baseline**	**mean rank on biased**
medicare	1	2	107.82	119.09
other	2	1	111.24	93.97
medicaid	3	3	143.75	155.04

Table 6.8: The ranking performed by FairLens using disparity score for the baseline and biased DSS.

We also compared FairLens average ranking aggregated by insurance type for both the biased and the baseline DSS. The results reported in Table 6.8 show that, for the baseline DSS, FairLens ranks *Medicare* as the insurance having the highest disparity score across different groups, while *Other* is ranked above the others for the biased DSS.

Finally, we measured the outcome disparity for the insurance variable using the multi-label standard metrics used to evaluate DoctorAI performance in the original paper, *recall@k* and the *microAUC*. We compared the difference of these metrics in the baseline

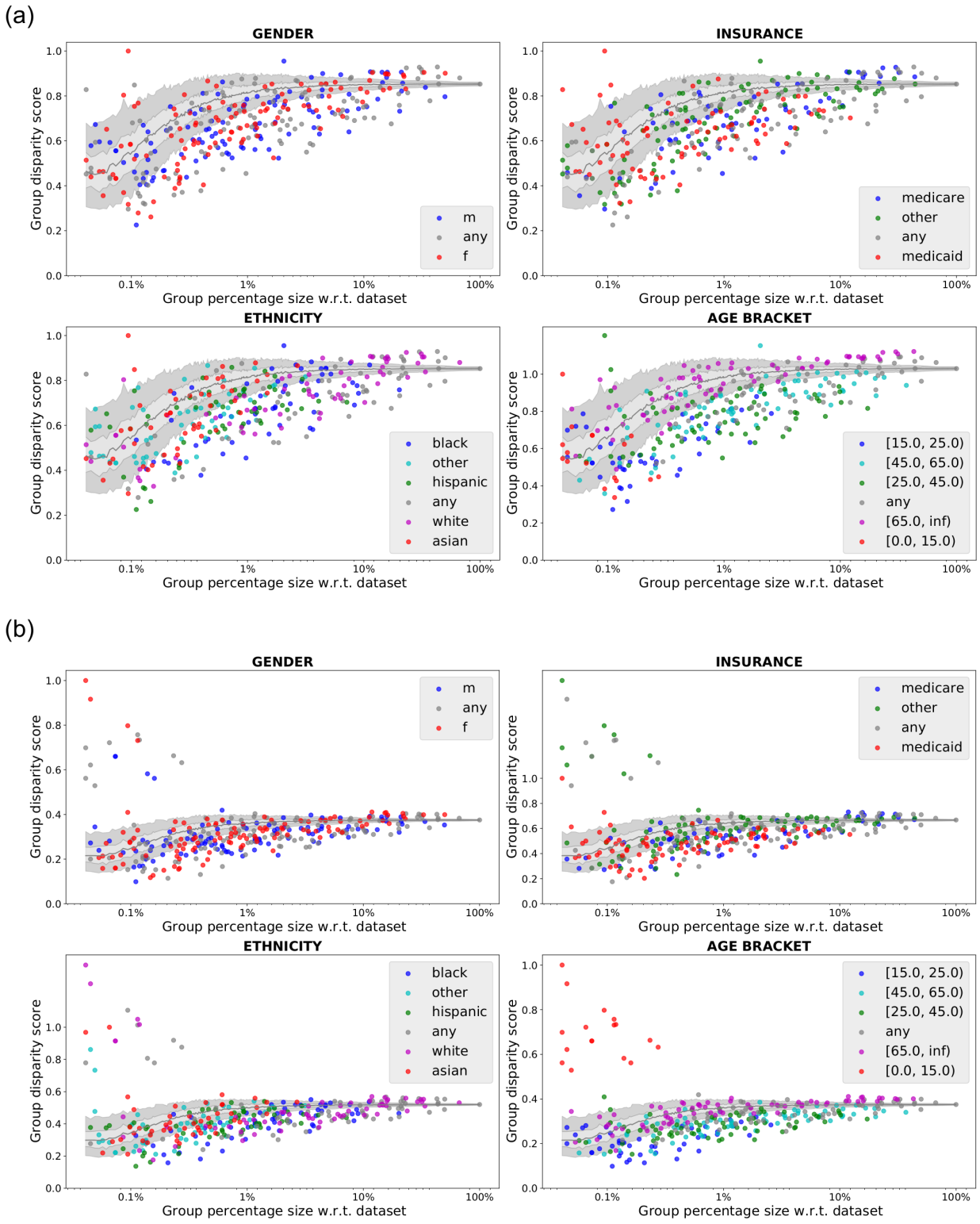


Figure 6-7: FairLens scatterplots for the baseline DSS (a) and biased DSS (b)

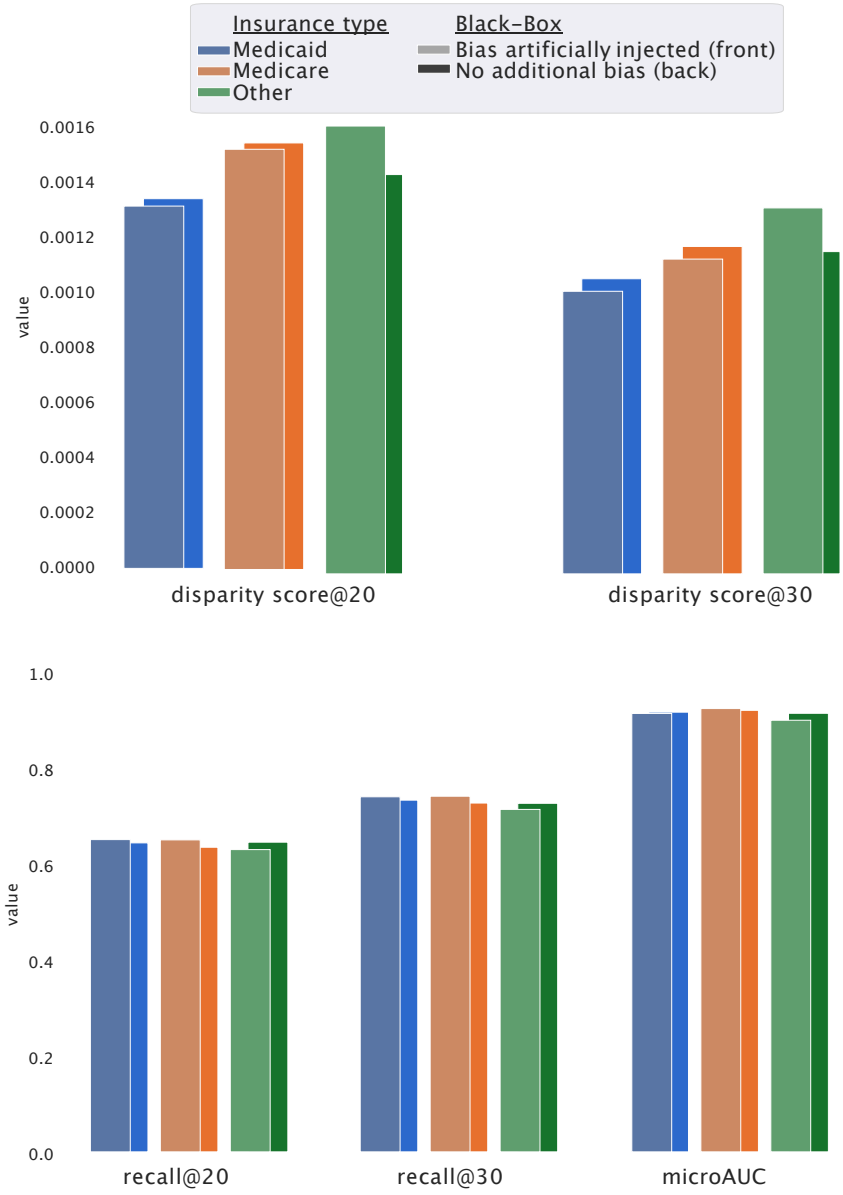


Figure 6-8: Average metrics across insurance groups

and biased DSS in Figure 6-8. We can see that while the standard metrics remain almost constant or slightly decrease in the biased BB with respect to the baseline DSS, both disparity scores evaluated at top $k = 20$ and $k = 30$ exhibit a clear increase for the under sampled group in the BB where bias was artificially injected.

6.7 Discussion

In this chapter, we explored the ability of XAI to audit a commercial black box. As also highlighted in chapter 4.5, systematic auditing procedures must be in place when black-box ML-based clinical DSSs are deployed in real-world healthcare settings.

It is essential to build external algorithmic auditing tools that allow an objective evaluation of the effectiveness and fairness of algorithmic systems. Even though an internal algorithmic auditing process is of pivotal importance to release a product that meets the ethical and reliability standards of those who developed and marketed the product, its cost-benefit analysis might be skewed toward maximizing profit. External auditing tools allow companies to be held accountable to third parties and increase the credibility of the algorithmic pipeline. Independent auditing is also helpful to test the model on the target population where the DSS should be deployed.

We proposed FairLens, an algorithmic pipeline to inspect clinical DSSs to spot potential fairness issues in patients' groups that call for further investigation of possible over-/under-diagnosed conditions. The proposed methodology can help domain experts investigate the reason behind the systematic black-box misclassification by pointing to the most common causes of error within groups through XAI techniques.

The main use-case presented in this chapter describes the auditing process of a clinical DSS trained on sequential visits to predict the diagnoses associated with the next patient's visit. FairLens can be generalized to other use-cases with different DSS tasks, as far as the building blocks are adequately adapted. The application of FairLens to a different clinical DSS is presented in Appendix A. While the final aim (auditing a black-box) and the intended user (IT expert responsible for deploying the DSS in the healthcare facility) are the same, the machine learning model is completely different. The experiment highlights the flexibility of our framework, adapted to work on predicting the ICD9 codes given the raw text of clinical notes, relaxing the temporal dimension of sequential visits. While the scoring mechanism remains unchanged, the explainability approach and the local-to-global aggregation mechanism are adapted to the prediction task.

DSS developers can also use FairLens to perform a sanity check of the model and detect and mitigate potential biases before its release. However, this would require ML engineers to know the medical domain or team up with medical personnel to understand if the potential bias signaled by FairLens reflects a real fairness issue.

It is worth stressing that FairLens is not designed to be an automated tool but rather to help human auditors in identifying groups where fairness issues may arise. Moreover, FairLens is not able to provide the origin of such misbehavior (e.g., eliciting if the source of bias is in the original training data, is embedded in the algorithm itself [126] or in the prediction task), as it is designed to perform external audit without having access to information about the black-box nor the original training data.

External auditing tools such as FairLens could also identify ICU patients' over/under-treatment to improve patient experience in the hospital. Under the assumption that high disparity scores suggest a mismatch between what the clinical DSS learned and how the patients were historically treated in the healthcare facility, the auditor might even find biases in the auditing data, leading to a quality assessment of hospital services.

It is also important to discuss potential uses of FairLens, which differ from the one envisioned and discussed in this chapter. Theoretically, if linked with information that leads to the identification of the operator responsible for patients' treatments, FairLens could be used to identify doctors that systematically treat groups differently. While doctor performance assessment is precious and several techniques to operationalize it already exist [265], such unintended use of FairLens should be properly considered.

While in this chapter, we shift our focus toward a real-world setting, the explanations provided by FairLens might be challenging for non-technical users. In the next chapter, we take a step back and focus on evaluating the effectiveness of Doctor XAI explanations (the explanations combined in a local-to-global fashion in FairLens) on healthcare providers.

Chapter 7

Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems

7.1 Introduction

The previous chapter focused on the ability of XAI to audit a fictional commercial black-box AI system whose software is proprietary. While the results show that this is possible from a technical point of view, we need to investigate whether the provided explanations are effective on real-world users. Indeed, up to this point, we did not include any human evaluation of the explanations provided by the presented methods. In this chapter, we take a step back and focus on evaluating Doctor XAI explanations' impact on a group of healthcare providers. Involving the end-users is of pivotal importance to evaluate if explanations can enable a real human oversight (objective 4 of chapter 3). Indeed, AI shows great potential for healthcare applications such as clinical Decision Support Systems (DSS). However, the rate of adoption of such technology in health clinics and hospitals is low [395, 13, 317]. A recent report estimated that 84% of healthcare providers in Europe currently do not use any AI system [159]. The reasons behind the low adoption of clinical DSS that do not embed AI have been well studied, and the difficulties include perceived

challenges to autonomy, lack of time, and dissatisfaction with user interfaces [208, 41, 246, 359, 311, 355, 196, 202]. In addition to these adoption barriers, AI technologies also face trust issues from medical staff and a lack of knowledge about their limitations and capabilities [159, 363]. Trust plays a central role in the adoption of new technologies and explanations of AI recommendations are often touted as the solution to trust issues [302, 114, 369, 392]. This chapter explores the relationship between trust, XAI explanations, and users' intention to adopt an AI-based clinical DSS. We studied these aspects by performing an online user study on the impact of AI explanations in the medical field. In the following we present the theoretical foundations of our experiment and related works.

7.1.1 Theoretical models of acceptance and use of technology

Human Computer Interaction (HCI) and psychology researchers have developed several theories to identify and explain the factors that predict users' acceptance and use of technology. More than one theoretical framework might be needed to study a particular issue. In our work, we followed the Technology Acceptance Model (TAM) [364] and the Unified Theory of Acceptance and Use of Technology Model (UTAUT) [366]. The central factor in these models is people's intention to perform a particular behavior, i.e., their *behavioral intention*. The behavioral intention is assumed to be the best predictor of actual behavior. Indeed, the greater the intention to engage in a behavior, the more likely its execution should be. These models propose different factors influencing behavioral intention to use technology.

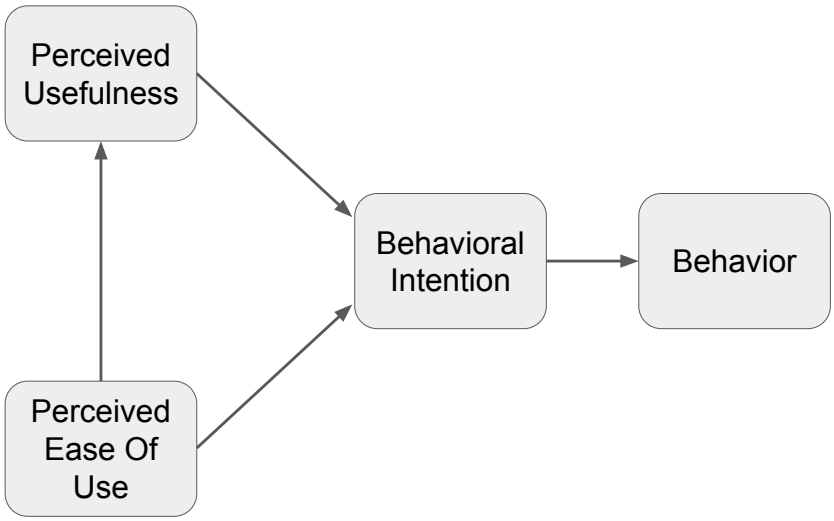


Figure 7-1: Technology Acceptance Model (TAM)

At its core, the TAM [92] (figure 7-1) and its subsequent extensions [365, 364] postulate that the two most important predictors of an individual's behavioral intention of using a certain technology are its *perceived usefulness*, defined as "the degree to which a person believes that using a particular system would enhance his or her job performance" and its *perceived ease of use* defined as "the degree to which a person believes that using a particular system would be free of effort".

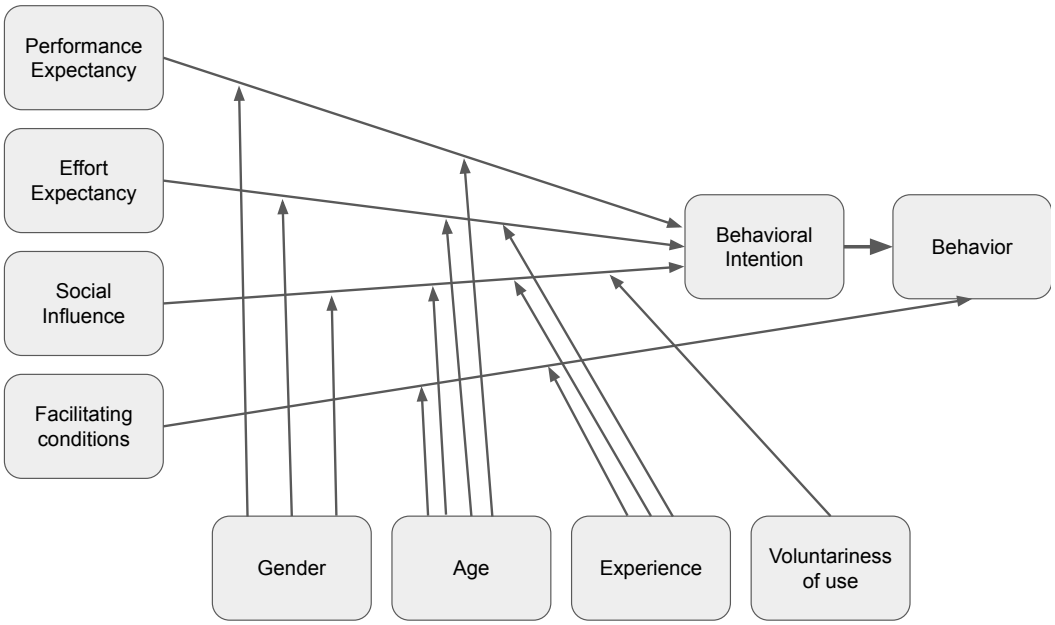


Figure 7-2: Unified Theory of Acceptance and Use of Technology Model (UTAUT)

The UTAUT (figure 7-2) integrates eight models of technology acceptance by identifying four key categories of direct determinants of behavioral intention. Some of these are in common with TAM. For example, it considers the *performance expectancy* and the *effort expectancy* constructs which can be respectively mapped to the *perceived usefulness* and the *perceived ease of use*. However, it also considers some organizational aspects. Indeed it also identifies as key constructs the *social influence*, defined as "the degree to which an individual perceives that important others believe he or she should use the new system" and the *facilitating conditions*, defined as "the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system". In the original UTAUT paper, gender, age, experience, and voluntariness of use have a moderating role on the relationship between the model's key constructs and behavioral intention. The UTAUT questionnaire has already been used in the healthcare setting to evaluate the factors influencing healthcare professionals' adoption of mobile electronic

medical record [198, 367] and to investigate the adoption of AI-based medical diagnosis support system [114].

7.1.2 Trust

Neither the TAM nor the UTAUT explicitly acknowledges trust as a construct of the model. Trust has been overlooked in these models because it was assumed to be associated only with interpersonal relationships [165, 121]. However, many studies have shown that humans respond socially to complex technology [254], i.e., when the system used goes beyond a simple tool with clearly determined and easily understood functions [210, 162]. Therefore, trust plays a central role in the adoption of AI-based technologies. We adopt the following definition of trust:

> Trust is the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid [234].

Trust in AI becomes vital in scenarios that entail risk, uncertainty, and vulnerability to negative outcomes [10]. For example, trust is not an essential factor in adopting an AI algorithm that ranks the documents in a folder according to their relevance, whereas trust is essential in a clinical decision-making context. Trust can be measured both by employing *explicit* and *implicit* measures. Explicit measures involve the use of trust scales that directly ask users whether they trust the AI or not [166], while implicit measures rely on operationalizing the definition of trust in terms of user behavior: does the user changes his or her behavior after receiving the AI-system suggestion? [396] Indeed, in the context of decision-making, trust is positively associated with advice taking [130, 336]. Advice taking can be measured using the *Weight Of Advice* (WOA) [156], i.e., the extent to which participants change their initial estimate after receiving the AI system's suggestion:

$$WOA = \frac{|F - I|}{|A - I|}$$

where I and F are respectively the pre- and post-advice judgments and A is the received advice. The WOA is often used with the Judge-Advisor System (JAS) framework [335, 336]. In a JAS there are two distinct roles in the decision-making process: the judge and the advisor. While the advisor provides to the judge suggestions and advice, the judge is the

only responsible for the final decision. In the context of clinical decision-making the clinical DSS is the advisor and the clinician is the judge, solely responsible to provide appropriate care for the patient.

7.1.3 Trust calibration and AI explanations

Ideally, explaining clinical DSS recommendations should help clinicians with *trust calibration*, i.e., to properly adjust their level of trust according to the actual reliability of the AI system [313]. There are several levels of trust falling along a spectrum ranging from complete distrust to overreliance on AI. Both extremes have been observed towards AI-based clinical DSSs. On the one hand, some works have shown that clinicians tend to over-rely on automated suggestions by taking less initiative [214] or accepting incorrect diagnoses suggested by AI [151]. This phenomenon is known as *automation bias* [210, 333] and can be particularly dangerous in critical domains such as medicine. On the other hand, physicians are reluctant to trust algorithms that they do not understand [56, 331] and might be subject to *algorithm aversion* [95], which is the human tendency to discount algorithmic advice [225]. Distrust in AI applications in medicine also comes from doctors' fear of legal repercussions if something goes wrong due to unclear liability regimes [255, 339]. While, at first glance, explanations of such DSS seem the solution to these issues, some studies suggested that explanations can be inadequate to deal with overreliance on flawed algorithms [174]. Furthermore, explanations might even increase overreliance on AI-based clinical DSS [53, 204], and it might be necessary to design the system to force the user to engage in analytical thinking when explanations require substantial cognitive effort to be evaluated [51]. These findings highlight the importance of involving the end-user of the explanation when evaluating its efficacy and, ideally, in the design phase. However, there is a gap between state-of-the-art XAI explanations and end-users. A recent survey has shown that machine learning engineers mainly use explanations of black-box AI systems to debug their model in the development phase [39], i.e., developers of XAI methods design explanations for themselves. In the medical field, a few works have tried to close such a gap by involving the doctors in the design procedure [392, 316, 205] or by performing exploratory surveys on their needs [352, 56, 220]. Despite these recent efforts, most of the research has been focused on laypeople [19, 248, 68]. However, several works have shown that users' domain expertise is relevant to the trust calibration process [258, 377, 412], e.g., novice users tend to over-rely on AI suggestions. For these reasons, in our study we focus

on the impact of explanation on advice taking involving a specific pool of end-users, i.e., healthcare providers, and observing the use of explanation in the appropriate decisional context [40, 237], i.e., while performing a task supported by a clinical DSS.

Finally, another important factor to consider is the perceived *explanation quality*. Indeed, *good* explanations enable end-users to develop an appropriate mental model of how the AI system works, facilitating the trust calibration process. To measure explanations quality, we employed the *explanation satisfaction scale* [166] which measures explanations' understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness from users' point of view.

7.2 Main contribution

In this chapter, we present the results of an online user study on the impact of AI explanations in the medical field. For our experiment, we considered our XAI methodology, *Doctor XAI*(presented in section 5.4), and we employed *Doctor AI* [72], a recurrent neural network, to act as clinical DSS. The purpose of our research is to understand how explanations could enhance the trust in the AI system and the intention of using an AI system in the medical field. Our research questions are the following:

- **RQ1**: How AI explanations impact users' trust in algorithmic recommendations in the healthcare context?
- **RQ2**: How AI explanations impact users' behavioral intention of using the system in the healthcare context?

In particular, we want to test the following main hypotheses:

- **Hp1**: Participants trust more the algorithmic suggestion when it is presented with the explanation.
- **Hp2**: Participants feel more confident when they use the system that provides an explanation
- **Hp3**: Participants have a higher behavioral intention to use the system that provides an explanation.
- **Hp4**: Participants express higher trust for the system with the explanation

7.3 Methods

7.3.1 Participants

We ran an online experiment on the *Prolific* platform (www.prolific.co). We prescreened participants to be healthcare providers (doctors, nurses, paramedics, and emergency services providers), fluent in English, and high acceptance rate. All participants provided written informed consent and studies were approved by local Research Ethics Committees. Each participant was asked to perform a task (detailed below) and answer a set of questionnaires and received a compensation of 6.20£ for it.

7.3.2 Estimation task

To evaluate whether the explanation of the algorithmic recommendation influenced participants' behavioral intention and trust in the clinical DSS, we used an *estimation task*. During the estimation task, the participant is asked to make an estimate before and after being presented with the algorithmic recommendation. In this case, the task was to estimate the chances of a patient suffering from an acute myocardial infarction (acute MI) in the near future. Participants were first presented with the patient's clinical history and asked to make an initial estimate based on their knowledge and experience. Then they were shown the algorithmic suggestion, and they were asked to make a second and final estimate (more details can be found in Appendix B.9). This task allowed participants to decide how much they want to rely on the algorithmic suggestion, weighing it compared to their first estimate. Our paradigm adapts to the judge-advisor system (JAS) [335, 336]

7.3.3 Experimental design

The experimental design followed a two-cell (only AI suggestion vs. AI suggestion and explanation) within-subjects design. Each participant was asked to perform the estimation task twice: once using the interface providing only the AI suggestion (blue path of figure 7-3) and once using the interface providing the suggestion and the explanation (yellow path of figure 7-3). To prevent the learning effect, each participant used the two interfaces on two different yet analogous patients. To prevent order effect, participants were randomly assigned to different experimental groups to control the order of presentation of the different types of algorithmic suggestions (with or without explanation).

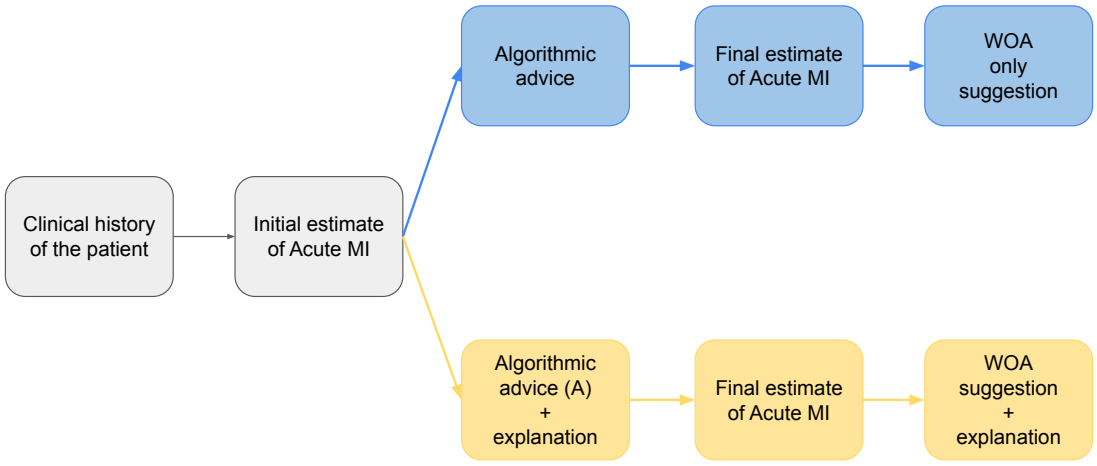


Figure 7-3: Flowchart of the estimation task for the two interfaces: only suggestion (blue path) and suggestion and explanation (yellow path)

7.3.4 Collected data

Implicit trust and confidence

Our main dependent variable was the Weight of Advice (WOA) [156] defined as follows:

$$WOA = \frac{|F - I|}{|A - I|}$$

where F and I are respectively the final and initial participant's estimates, while A is the algorithmic suggestion. Participants were asked to estimate the patient's chances of developing an acute MI in the near future on scale from 0 to 100% and their confidence in the estimate on a sliding scale. To avoid adding further degrees of freedom to the experiment, we selected only patients correctly predicted by the algorithm as having an acute MI in the near future, therefore $A = 100$ in all cases. Participants were also asked to indicate their *confidence level* after each estimate.

Explicit trust

In addition to the WOA, we also measured the *explicit* trust in the system by directly asking participants' perception on the system reliability, predictability, and efficiency (5-point Likert scale, from 1="strongly disagree" to 5="strongly agree") [166, 55, 10].

Behavioral intention and correlated constructs

To measure and compare the Behavioral Intention (BI) of using the two interfaces, we adapted the UTAUT and the TAM questionnaires from [364, 366] (the whole questionnaire can be found in Appendix B.11). In particular, we collected the following constructs (5-point Likert scale, from 1="strongly disagree to 5="strongly agree"):

- **Performance Expectancy**: the degree to which an individual believes that using the system will help him or her to attain gains in job performance [366].
- **Effort Expectancy**: the degree of ease associated with the use of the system [366].
- **Attitude Towards using Technology**: an individual's overall affective reaction to using a system [366].
- **Social Influence**: the degree to which an individual perceives that important others believe he or she should use the new system [366].
- **Facilitating Conditions**: the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system [366].
- **Image**: the degree to which use of an innovation is perceived to enhance one's image or status in one's social system [364].
- **Job relevance**: The degree to which an individual believes that the target system is applicable to his or her job [364].
- **Output Quality**: The degree to which an individual believes that the system performs his or her job tasks well [364].
- **Result Demonstrability**: The degree to which an individual believes that the results of using a system are tangible, observable, and communicable [364].

Explanation satisfaction

We measured the perceived explanation quality using the explanation satisfaction scale (5-point Likert scale, from 1="strongly disagree to 5="strongly agree") proposed in [166] and collected qualitative feedback using open-ended question on participants' experience using the two AI interfaces (see Appendix B.15 for the complete list of questions).

Confounding factors

We controlled for confounding factors such as participants' familiarity and involvement in the task [99], demographic information such as gender, age, and the type of medical profession (see Appendix B.5). We also controlled for participants' Need For Cognition (NFC) - an aspect related to the individual tendency to enjoy effortful cognitive tasks (5-point Likert scale, from 1="strongly disagree to 5="strongly agree") [54, 239] (see Appendix B.4). We now proceed to illustrate the two AI interfaces used in our experiment.

7.3.5 Interface "Only suggestion"- Dr. AI

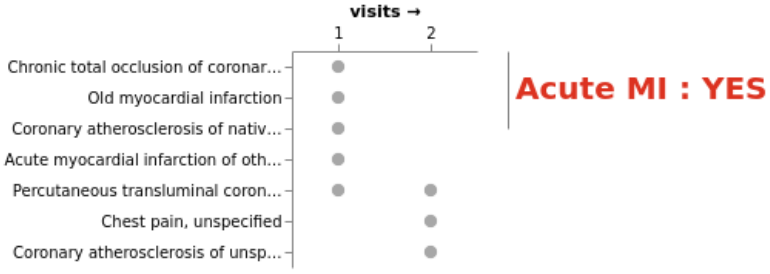


Figure 7-4: Static visualization of the *only suggestion* AI interface

Acting as clinical DSS, we used Doctor AI [72], a Recurrent Neural Network able to predict patients' future diagnoses based on their past clinical histories. We post-processed Doctor AI outcomes transforming them from multi-label (every diagnosis of future visits) to binary to predict whether a patient would have an acute MI or not. A static visualization of the interface providing only Doctor AI suggestions is shown in figure 7-4. The visits of the patients are represented as a set of grey dots, and each dot represents a condition diagnosed in the corresponding visit. For example, this patient was diagnosed with five conditions in their first visit and with three conditions in the second one. In the dynamic visualization, participants were able to explore the conditions diagnosed in each visit and visualize their descriptions by moving the cursor over the corresponding dots. Finally, the AI suggestion is shown in red to the left of the patient's clinical history.

7.3.6 Interface "Suggestion and explanation" - Dr. XAI

To extract an explanation for the algorithmic suggestion, we employed Doctor XAI [271], an eXplainable AI (XAI) technique able to deal with sequential clinical histories that

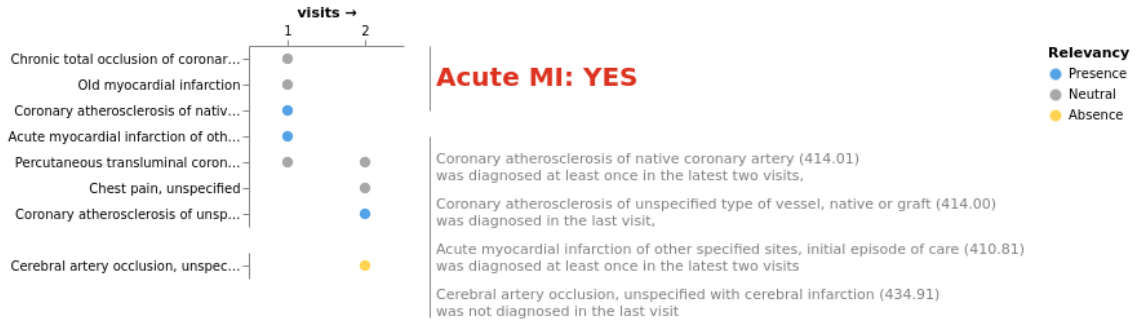


Figure 7-5: Static visualization of the *suggestion and explanation* AI interface

use medical knowledge in its explanation extraction process. Doctor XAI's explanations highlight which conditions in the clinical history of the patients were deemed the most important ones by the algorithm in its decision-making process. Furthermore, Doctor XAI also provides information regarding the missing conditions that influenced the algorithmic decision. A static visualization of the interface providing AI suggestions and explanations is shown in figure 7-5. Doctor XAI assigns a different color to each dot according to the corresponding condition's relevance to the algorithmic decisions. Dots corresponding to conditions deemed irrelevant are left grey, while dots deemed relevant are colored blue. Furthermore, Doctor XAI shows as yellow dots conditions that are missing from the patient's clinical history that would have changed algorithmic suggestion. Finally, a summary of the explanation is written under the algorithmic suggestion. The dynamic visualization allowed participants to highlight the conditions in the clinical history corresponding to each sentence in the written explanation summary.

7.4 Results

7.4.1 Quantitative analysis

A total of 31 healthcare providers participated in the online experiment. The analysis discarded three participants: one did not pass the attention check question, while two gave 100 as their initial estimate, which yielded undefined values for the WOA ($A = I$). Eventually, 28 participants were retained for the study. 5 doctors, 20 nurses, one health care assistant, one dietetic assistant practitioner, and one ambulance call dispatcher. The mean age was 41 years old (SD=11) ranging from 24 to 65 years old. 21 were women and 7 men. The male sample has a mean age of 34 years old (SD=9), and the female sample

has mean 43 years old (SD=11). We performed all the analysis in Python.

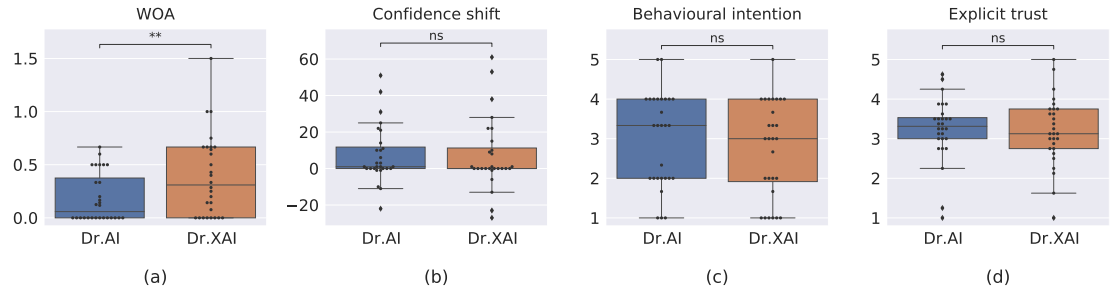


Figure 7-6: Boxplot comparing the WOA (a) the confidence shift after the advice (b) the behavioural intention of use and (c) the explicit trust in the two systems (d).

Weight of Advice and Confidence. In figure 7-6(a) we show the result of the comparison between the WOA for the two AI interfaces: Dr.AI (only suggestion) and Dr.XAI (suggestion and explanation). The WOA was higher for the Dr.XAI interface (Mdn=0.31) than the Dr.AI interface (Mdn=0). A paired-samples two-sided Wilcoxon signed-rank test indicated that this difference was statistically significant ($T = 32.5$, $p = 0.002$). This confirmed our first hypothesis showing that participants were more influenced by the AI interface showing an explanation for its recommendation. Since advice-taking is positively correlated with trust, we can interpret this result saying that, on average, participants implicitly trusted more the AI interface that provides explanations. In figure 7-6(b) we compared participants' confidence shift for the two interfaces. The confidence shift was measured as the difference of the reported participant's confidence in the estimate before and after receiving the AI advice. A paired-samples two-sided Wilcoxon signed-rank test did not find any statically significant difference between the two interfaces $T = 169$, $p = 0.869$. This means that the explanation did not significantly increases or decreased participants confidence in their second estimate compared with a system that provide only the suggestion.

Behavioural Intention and explicit trust In figure 7-6(c) we compared the behavioural intention of use for the two AI interfaces. A paired-samples two-sided Wilcoxon signed-rank test did not find any statically significant difference between the two interfaces $T = 37$, $p = 0.076$. This did not allow us to confirm our second hypothesis that the behavioural intention of use of the AI interface Dr.XAI (suggestion and explanation) was higher than

UTAUT variable	median Dr.AI	median Dr.XAI	Wilcoxon statistic	p-value
Performance Expectancy	3.2	3.0	66.0	0.391
Effort Expectancy	3.6	3.5	66.5	0.016(*)
Social Influence	3.2	3.5	74.5	0.403
Facilitating Conditions	3.2	3.5	79.5	0.333
Attitude toward techology use	3.2	3.1	100.0	0.587
Image	2.0	2.2	31.5	0.325
Relevance	3.7	3.3	40.5	0.726
Output quality	3.2	3.0	64.0	0.553
Result Demonstrability	3.8	3.8	128.0	0.224

Table 7.1: Comparison of UTAUT variables for the two interfaces. Median, paired sample Wilcoxon signed-rank test statistics and p-value.

the Dr.AI (only suggestion) one. However, our results also indicated a significant positive Spearman correlation between the behavioural intention of use of the Dr.XAI interface and the perceived explanation quality $rs(27) = 0.67$, $p < .001$. Similarly, we did not find a significant difference in explicit trust between the two interfaces (figure 7-6(d), paired-samples two-sided Wilcoxon signed-rank test, $T = 157.0$, $p = 0.881$), but we found a strong positive Spearman correlation between explicit trust and perceived explanation quality ($rs(27) = 0.77$, $p < .001$). This could indicate that this particular type of explanation does not suit healthcare providers well. Indeed, like those of most state-of-the-art XAI methods, such an explanation was developed and designed with debugging purposes in mind rather than to fit the specific needs of the final user. Therefore, healthcare providers perceive this explanation as unsatisfactory and do not increase their behavioral intention of use or trust in the system when presented with it.

Further findings In table 7.1 we show a comparison between the UTAUT variables in the two interfaces together with their medians and the related paired sample Wilcoxon signed-rank test statistics and its p-value. The only statistically significant difference between the two interfaces is *Effort Expectancy*. Given the small sample size, we leave to future works the creation of two models investigating which factors impact the most the behavioural intention. Furthermore, no statistically significant correlation between the confounding variables, the WOA, and the behavioural intention was found with Spearman correlation tests. The only relevant negative correlation was found between the WOA of the Dr.AI interface (only suggestion) and the single-item measure of familiarity with the task (rs(27)=-0.58, p-value = 0.001). This means that the algorithmic suggestion had a stronger influence on participants less familiar with estimating the chances of an acute MI. Finally, a Wilcoxon signed-rank test showed a slight difference in the WOA between

the different types of healthcare providers $T = 2.56$, $p = 0.025$. However, given the small sample for each category, we leave such an analysis for further works.

7.4.2 Open-ended questions insights

In order to evaluate participants' impressions, we asked them to answer open-ended questions. Participants' open-ended responses were coded through thematic coding [287]. Specifically, the analysis was carried out to create as few categories as possible without making them too broad.

Participants' perceptions and preferences Understanding users' preferences for one interface over the other is of pivotal importance to analyze their impressions. We asked the participants to give us answers about: 1)their general impression of each interface, 2) what they liked the most about the interface they had just used, 3)what they dislike the most about the interface they just used. Most participants appreciated the two interfaces, with slightly more participants leaving positive comments on the Dr.XAI interface (Dr.AI= 39.29%; Dr.XAI= 53.57%). Indeed, most participants did not appreciate the simple suggestion provided by the Dr.AI interface without any other information (54% of the participants asked for an explanation, while 46% did not express any opinion):

> *It is simple. Too simple in fact.* F, 36, Nurse
>
> *I wish this AI interface would provide more information about how it reached it's decision.* F, 40, Nurse.

However, when provided with the explanation, they were left unsatisfied by it:

> *Using the AI interface with the explanation built in was something I anticipated making the decision easier, but in fact this was not the case. All the information presented too much on the screen and took a lot of time to interpret and synthesise. Decision-making became more of a lengthy and arduous process.* F, 24, Doctor.
>
> *I think it has a lot of potential, but would like a more detailed rationale of why it thinks an MI is likely and a numeric assessment of how likely (as I was asked to give).* F, 51, Doctor.

Some suggested implementing a natural language version of the explanation and adding the time length between visits. Overall, participants did not encountered many difficulties (Dr.AI =85 %; Dr.XAI=68%). One of the common issues was understanding how to interact with the explanation. The explanation interface was considered useful to prevent novices from making mistakes and during collaborative decision-making tasks:

> *It would prevent novices making mistakes.* F, 52, Doctor.

> *The doctors in our acute medical department are very keen to discharge patients home; leaving nurses in a difficult predicament when we don't agree with their decision making. A tool such as this, could help nurses to justify their reasons for keeping a patient in hospital or to use cardiac monitoring vs. not monitoring.* F, 36, Nurse.

Algorithm aversion and fear of being replaced Eventually, one of the most surprising findings we came across is related to the participants' perceived threat of being replaced by the AI system. In both conditions, comments like the ones reported below were common:

> *Can be useful but does not replace human judgement.* F, 59, Nurse. (Dr.XAI condition).

> *it could be taken as fact that the AI is correct which disregards the human factor and individuality.* F, 53, Nurse. (Dr.AI condition)

> *It was really good but human health isn't always black and white. You can't put AI in human nature. Yes it may use stats probabilities etc but there's always that one patient that goes against the rules. I'd use it to as a tool to bear in mind but I wouldn't rely on it. [...] It takes away the thinking this the prestige of all the effort and study you've put in!.* F, 39, Nurse (Dr.XAI condition).

While this might be associated with the phenomenon of *algorithm aversion* [95], or the human discount of algorithmic advice [225], the prevailing sentiment emerging from such open-ended questions was the fear of being replaced by AI. This fear of being replaced is often an underestimated aspect in computer science research, however, the understanding of the sociocultural environment in which the user operates has a paramount relevance in the acceptance of such AI systems [102].

7.5 Discussion

This chapter examined whether XAI explanations enable human oversight in the clinical context. Indeed, AI explanations should help the physician establish the right amount of trust in the clinical decision support tool and help her understand when intervening is necessary.

We performed an online user study adopting the specific lens of the Weight of Advice (WOA), the Trust Scale, and the Behavioral Intention from the TAM model. We compared two interfaces for an AI-based clinical DSS by manipulating how the suggestion was presented to the healthcare providers (with or without explanation) and asked them to perform the estimation task before and after interacting with the two interfaces.

We found that participants were keener on taking advice from the AI interface that explained its suggestion than the one that did not. This was reflected in a greater shift in the estimates provided after receiving such algorithmic advice, i.e., the weight of advice.

We gain even more insight into the effect of the explanation from the open-ended questions. The answers suggested that participants did not appreciate the suggestion alone and preferred an explanation for it. However, the explanation provided left most of them unsatisfied.

It is interesting to notice that, despite the low perceived explanation quality, participants were influenced by it and relied more on the advice of the AI system. This finding might be in line with previous research on *automation bias* in medicine, i.e., the tendency to over-rely on automation [134, 164, 221], and might significantly impact the ability of explanations to ensure an appropriate level of human oversight.

We also studied the confidence after the advice and the explicit trust in the system, finding no significant differences between the two interfaces. Similarly, we did not find a significant difference in the behavioral intention (BI) of the use of the two interfaces. A possible explanation for it is the high correlation between the BI and the perceived explanation quality, i.e., the proposed explanation was not appropriate for the healthcare audience. However, from the open-ended questions emerged an alternative interpretation of this finding. Indeed, many participants showed some degree of algorithm aversion and expressed the fear of being replaced by the AI system. The AI system was perceived as threatening human judgment rather than as a decision support tool. This finding is relevant in the design of AI applications in healthcare regardless of XAI explanations and

shows that it is crucial to have an interdisciplinary approach to comprehend the factors that influence technology adoption.

This study has some limitations. First of all, a limited sample size. In future work, we aim to carry out a more complete and accurate study differentiating different healthcare providers' needs, also considering different task-related expertise.

Furthermore, Doctor XAI explanations are representative of a single type of explanation. However, Dr.XAI's explanations are both medical domain-aware and a good representative of a common type of AI explanation: the *removal-based type* of explanation [85]. Like other popular removal-based approaches, Dr.XAI explanations summarize each feature's influence on the model outcome [302, 228]. However, unlike other removal-based approaches, it also employs medical knowledge in the explanation extraction process, meaning that the features highlighted to be important were selected considering their medical meaning. These explanation characteristics are therefore well suited for our purpose of evaluating the impact of AI explanations on healthcare providers.

Chapter 8

Final discussion and future work

In this thesis, we explored the ability of XAI techniques to meet different requirements for trustworthy AI in the context of healthcare applications. Chapter 4 analyzed the legal and ethical framework related to the development and use of AI systems in healthcare. The chapter provided an overview of the different approaches to AI ethics worldwide and mapped some ethical values to the various stages of the ML development lifecycle. Then the analysis focused on the EU legal and ethical framework and identified three main trustworthy AI requirements relevant to XAI: *transparency*, *auditability*, and *human oversight*. Each chapter of this thesis focused on one of these requirements. However, each of these principles is related to the others. For example, to ensure human oversight, it is necessary to have transparency and auditability. It is important to note that these principles are necessary but not sufficient for trustworthy AI. Our analysis focused on this specific subset of requirements because XAI techniques can help ensure their fulfillment. The following sections discuss how the work presented in the thesis is an appropriate interpretation of these high-level values and how it supports their operationalization.

8.1 Transparency

.

The EU ethics guidelines for trustworthy AI [163] state that the transparency requirement is linked with the principle of *explicability* and *"encompasses transparency of elements relevant to an AI system: the data, the system, and the business models"*. In particular, this requirement is broken down into three sub-requirements: *traceability*, i.e., the docu-

mentation of all operations related to data and AI design choices; *communication*, i.e., the communication of the limitations and capabilities of the AI system to the stakeholders; and finally *explainability*, i.e., the ability to explain the technical process of the AI system and how the human decision-maker interacts with it. While the traceability and communication sub-requirements can be achieved by adequately operationalizing record-keeping and establishing appropriate communication procedures, achieving technical explainability for black-box AI systems requires developing XAI techniques.

This thesis focused on developing new solutions to the *outcome explanation problem* (section 5.1.1), i.e., the problem of providing an explanation for a specific black box outcome. This is problem is explicitly mentioned in the EU guidelines for trustworthy AI when the *principle of explicability* is discussed as one of the fundamental ethical principles in the context of AI systems. In particular, the document mentions its interpretation of explainable AI decision as "*an explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that)*". Adopting such an interpretation, in chapter 5, we presented two novel XAI techniques to explain the decision-making process that an AI system employs when making a prediction. In more technical terms, the chapter adopted a *model-agnostic* and *local* approach to XAI. Since the main focus of this thesis is on AI in healthcare, we focused on developing XAI solutions to the outcome explanation problem tailored to healthcare data characteristics such as multi-label outcomes and sequential and ontology-linked data.

Chapter 5 focused on evaluating the quality of explanations from a technical point of view, i.e., using metrics of explanation goodness such as fidelity (i.e., whether the explanation faithfully captures the decision-making process of the underlying black-box model) and explanation complexity (i.e., the length of the explanation rule). While these two metrics are respectively proxies of the *trustworthiness* and *comprehensibility* of an explanation, they do not take into consideration how the human decision-maker interacts with it. We investigated this aspect in chapter 7 that focuses on the issue of *Human oversight* (section 8.3).

8.2 Auditability

The EU ethics guidelines for trustworthy AI [163] state that *"auditability entails the enablement of the assessment of algorithms, data and design processes"*. This requirement needs to be interpreted considering its interrelationships with other trustworthy AI requirements such as the *diversity, non-discrimination and fairness* one (see figure 4-1). Indeed, algorithmic auditing should be aimed at ensuring that the AI system decision making is fair and is not based on biased data. XAI can enable the assessment of black-box algorithms by uncovering their decision-making process and allowing an external auditor to inspect whether such a process is fair and unbiased. This thesis explored this topic in chapter 6, where we focused on using the XAI methodology presented in chapter 5, called *Doctor XAI*, to audit a black-box clinical decision support system and explain its systematic biases on selected groups of patients. We validated our framework, called *FairLens*, by injecting synthetic bias in the training set of the black box and proving that our methodology was able to detect it. We also proved the feasibility of using XAI for external auditing whose goal is to identify fairness issues due to data biases. Indeed, once a particular AI misdiagnosis is identified as systematic for a certain group of patients (e.g., patients belonging to a legally recognized group of patients), FairLens exploited Doctor XAI, the local and model-agnostic XAI technique of chapter 5, to uncover the black box decision-making process valid for that group of patients.

Two other aspects are important to notice. First, FairLens allows to tackle *intersectional bias*, i.e. bias that affect people based on the combination of several aspects of their life (e.g. gender, race, socio-economic status). Indeed, it allows to evaluate the disparity score of groups having any combination of protected attributes. This is in line with the EU's digital strategy *"A Europe fit for the digital age"* [81]. Second, given the fact that the definition of fairness is highly context-dependent (e.g. it depends on societal values), FairLens leaves the final decision on whether the observed bias constitutes a fairness issue to the external auditor. This is possible because the disparity score of FairLens is accompanied by an explanation that allows the external auditor to inspect whether the reasons behind the systemic misdiagnosis represents a fairness issue.

8.3 Human oversight

The EU ethics guidelines for trustworthy AI [163] state that *"Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects"*. As indicated in the guidelines, human oversight might be achieved through different governance mechanisms involving a human overseeing the AI system. However, to enable human oversight over AI systems it is necessary for the human to have an understanding of how the system works and performs its prediction. XAI can enable human oversight by providing such insights into the AI model decision-making process. An interdisciplinary approach to XAI is fundamental to study AI explanations effectiveness in enabling human oversight. Indeed, the goodness of an explanation does not lie in the XAI method but in the perceptions of the person receiving the explanation. This aspect is clearly illustrated in article 13 and 14 of the AIA. Indeed, article 13 prescribes that and high-risk AI system should be designed to allow the user to appropriately interpret its output. Furthermore, article 14 prescribes the design of appropriate human-computer interfaces to allow human oversight.

We investigated this topic with a user study in chapter 7. Indeed, to achieve proper human oversight over AI-based clinical decision support systems is essential to study how the related explanations impact advice-taking. AI explanations play a central role by allowing the healthcare provider to inspect the factors that led the AI system to make a particular recommendation, potentially preventing its mistakes, i.e. overseeing the system. Building adequate AI explanation interfaces is the fundamental step toward allowing humans to oversee high-risk AI systems effectively. However, our user study results suggest that participants were keener on taking the advice when presented with an explanation, even if it was perceived as a low-quality one. This finding insinuates doubt that AI explanations might increase the risk of automation bias, i.e., the tendency to over-rely on automation. Our findings align with previous studies that suggested that AI explanations can be inadequate to deal with overreliance on flawed algorithms [174], and that they might even increase overreliance on AI-based clinical DSS [**?** 53, 204]. To prevent overreliance on AI systems might be necessary to design the system to force the user to engage in analytical thinking when explanations require substantial cognitive effort to be evaluated [51].

8.4 Future work

The chapters of this thesis progressively broadened the view from purely technical to socio-technical and human-centered. Future work will be devoted even more to study the human aspects of XAI. The starting point is provided by the two interesting elements emerged from the user study of chapter 7: the tendency of AI explanations to increase automation bias, i.e., the tendency to over-rely on automation, and healthcare providers' algorithm aversion and fear of being replaced by AI. Future work will explore the impact of explanation on automation bias in medicine by focusing on a different pool of healthcare providers (experts in the predictive task) and considering wrong algorithmic suggestions. We will also further study healthcare providers' algorithm aversion using fictional scenarios and focus groups to better understand the factors influencing this tendency.

Another aspect that will be investigated is how to improve the explanation interface adopting a human-centered co-design process. Co-design is a participatory design approach that involves the end users as active participants in the design process [288]. This approach to design is particularly useful when developing human-centered XAI interfaces whose goal is to facilitate smooth and useful human-machine interactions. The objective of human-centered design is to "*enable users to achieve goals effectively, efficiently and with satisfaction, taking account of the context of use*" [172]. In the context of clinical DSS, a human-centered design of their user interface should increase healthcare providers' ability to understand and correctly act upon the DSS suggestions, which ultimately should result in improved health outcomes for their patients.

Finally, from a more technical point of view, we will investigate how to create trustworthy explanations of the global behavior of black-box DSS starting from the local explanations of Doctor XAI. Ideally, this process would allow to create a fully transparent model starting from a black-box. This promising line of research is in its infancy. One approach to solve this *local to global* [325] problem was used in the FairLens framework to merge the local explanations of the misdiagnoses of a group of patients into one explanation valid for all of them. In this sense, FairLens can be considered the first step in this direction.

The key takeaway of this thesis is that explainability is not just a technical issue, espe-

cially in the healthcare context. While technical transparency might be the first necessary step toward understanding black-box AI systems, a truly interdisciplinary approach involving medical, legal, and HCI experts is needed to reach the final goal of trustworthy AI in medicine.

A

Appendix A: FairLens use case 2 - auditing a clinical DSS for predicting medical codes from clinical notes

This section shows how FairLens can be used to audit a medical DSS that supports a user on the assignment of ICD-9 medical codes to a patient discharge, assuming that there are clinical notes associated to that patient. Convolutional Attention for Multi-Label classification (CAML) [249] is a medical decision support system that predicts medical codes from clinical text. As described in section 6.5, we imagine that the domain expert has no access to the source code of the DSS, i.e. it can be considered a *black-box*.

We use the MIMIC-III (see Subsection A.1) database of electronic health records as the fictitious historical database of the facility. The explainer in this use case is a model-aware attention mechanism, since CAML implements an attentional convolutional network that uses the attention mechanism to identify meaningful explanations for each code assignment. We split the dataset exactly as described in CAML paper, and we use the pre-trained CAML model trained to predict the MIMIC-III full set of 8.922 ICD-9 codes. For the pre-trianed model, training and validation set were used to train the CAML model. We use the samples in the test set as the historical database for the auditing process.

A.1 Dataset: MIMIC-III

MIMIC-III is the third update of the MIMIC (Medical Information Mart for Intensive Care) [135, 188] database. MIMIC-III is a single-center freely available database containing de-identified clinical data of over 40.000 patients admitted to the ICU (intensive care unit) of the Beth Israel Deaconess Medical Center in Boston collected from 2001 to 2012. The dataset contains patients' demographics, clinical measurement, billing information, medical history and also clinical notes written by healthcare providers in free-text format together with ICD classification for every visit.

A.2 Clinical DSS: CAML

The Convolutional Attention for Multi-Label classification model (CAML) [249] implements an attentional convolutional network to predict ICD-9 medical codes from clinical notes. Pre-trained models are available in many flavors: it can be trained on different versions of MIMIC (MIMIC-II or MIMIC-III), it might predict ICD-9 medical codes among the top 50 most common codes or among the whole set of codes, and it might include a regularization term to encourage each code's parameters to be similar to those of codes with similar textual descriptions, dubbed Description Regularized CAML (DR-CAML).

In this experiment, we used the non-regularized CAML model, pre-trained on MIMIC-III for the prediction of ICD-9 medical codes among the full set of 8.922 codes. We opted for the non-regularized version because it shows better performance for this specific setting. The source code for the pre-processing of the MIMIC-III dataset, for the train-validation-test split and the pre-trained models are available on the original paper's Github repository[1].

A.3 Local Explainer: The Attention Mechanism

While in the FairLens framework a post-hoc explainer that is agnostic w.r.t. the machine learning model would be preferable to ensure an auditing process that is fully disjoint by the model development phase, in this use case we leverage the attention mechanism as a XAI technique to explain the model outputs. The attention mechanism falls into the class of model-aware XAI techniques, since the attention scores produced by the model are used

[1] https://github.com/jamesmullenbach/caml-mimic

to measure how much an input feature explains the model output. In CAML, a convolution layer with filter size $k = 10$ is at the input of the attention layer, thus an attention score is assigned to every 10 consecutive tokens in a rolling window over the input. Therefore, by summing the scores of each token, the explainer is a function:

$$\xi : (\text{BB}, x_i, \beta) \rightarrow \{\ t_1 = s_1, \cdots, t_k = s_k\}$$

where BB is the CAML model, x_i is the the list of tokens extracted from the clinical note i, β is the ICD-9 code from which the output has to be explained. The function produces a set of explanations that assigns a score s_k to each token t_k.

Notice that attention scores are the output of a softmax operation, i.e., strictly positive values between 0 and 1, and thus have limited use as explanation scores [176], since they do not include text fragments that have a negative impact in the model decision. While such attention scores can be transformed in values that can be mapped to explanations when properly weighted, as shown in [44], the explanations produced by CAML do not use such approach.

A.4 Local-to-global approach

In the same spirit as GlocalX, the local-to-global approach used in this example aggregates local explanations of the form $\{\ t_1 = s_1, \cdots, t_k = s_k\}$ to a global set of rules $\{\ t_1 >= s_1, \cdots, t_k >= s_k\}$ that includes all explanations for the cases in which BB misclassifies an ICD-9 code β. Due to the structure of the explanations produced by the explainer in this use case, a simple local-to-global approach is to associate each term t_k to a rule $t_k >= s_k$ where s_k is the minimum score found in the set of explanations.

A.5 Auditing CAML on MIMIC-III

For this use case, only patient visits with clinical notes associated to them were selected from the MIMIC-III dataset. In particular, by following the same data preparation described in CAML, we select only discharge summaries and their addenda. The dataset contains 52.726 of such summaries, from which 49.354 were used for training and validation. We use the remaining 3.372 test entries as auditing data. Finally, there are 8.921 unique ICD-9 codes to be predicted by the model.

A.5.1 Assessing the DSS performance on the healthcare facility data

The performance of CAML on the auditing data is exactly as reported in the article [249], since the auditing data is the same used to test the model:

	@8	@15
Precision	0.709	0.561
Recall	0.373	0.526
F1	0.489	0.543

Table A.1: CAML performance on auditing data

A.5.2 Identifying problematic groups of patients

The following attributes were considered: *Gender*, *Ethnicity*, *Age* and *Insurance*. The scatter plots in Figure A-1 show the relationship between the group size and its normalized disparity measure. Each point is a combination of attributes, considering all possible permutations.

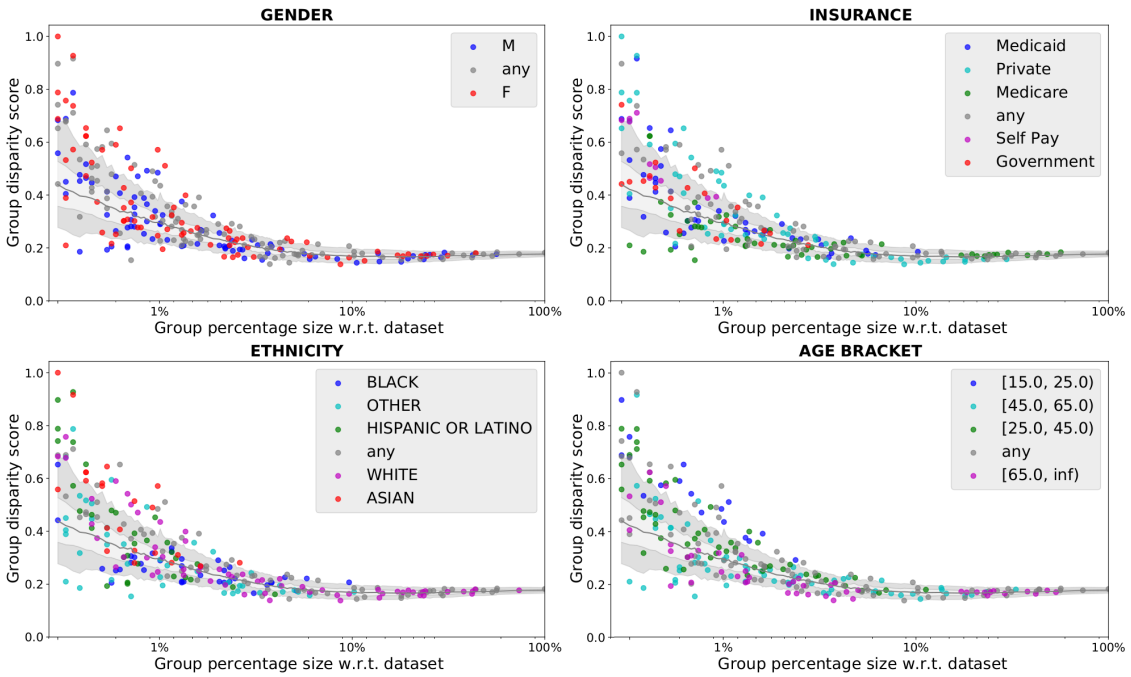


Figure A-1: Normalized disparity scores vs. group sizes with 50% and 95% bootstrap outliers bands.

Group	Size	Disp. Score	Over-diagnosed	(Misdiagnosis Score)	Under-diagnosed	(Misdiagnosis Score)
Female, Private, Asian	10	1.0	305.1: Tobacco use disorder	0.02	288.60: Leukocytosis	-0.037
			584.9: Acute kidney failure	0.013	872.02: Open wound of auditory canal	-0.033
			272.4: Other and unspec. hyperlipidemia	0.012	801.21: Closed fracture of base of skull	-0.029
15-25 years Private	54	0.39	96.6: Enteral infusion of conc. nutr. subst.	0.008	276.2: Acidosis	-0.01
			780.39: Other convulsions	0.006	275.3: Disorders of phosphorus metabolism	-0.008
			E950.4: Suicide and self-inflicted poisoning	0.006	807.01: Closed fracture of one rib	-0.007
45-65 years Medicare	224	0.21	428.0: Congestive heart failure	0.005	285.9: Anemia	-0.005
			38.93: Venous catheterization	0.004	458.29: Other iatrogenic hypotension	-0.004
			584.9: Acute kidney failure	0.004	V15.81: Personal history of noncompliance	-0.003

Table A.2: Groups ranked by normalized disparity scores for different group sizes and most over/under-diagnosed conditions when auditing the black-box

A.5.3 Identifying systematic sources of error in the selected subgroup

Table A.2 reports the the most over- and under-diagnosed CCS codes after computing the mistiagnosis score of each ICD9 code (we show only the 3 of the top groups by disparity scores for different population bins, and only the top/bottom codes ranked by misdiagnosis scores are reported). Due to the small number of examples included in the auditing data, the first two groups in the table are represented by a small number of patients (10 and 54), and might be discarded for being low represented. The third group, on the other hand, is represented by 224 samples.

A.5.4 Obtaining explanations for systematic misclassifications

For the last step of the pipeline, we focus on the largest group of Table A.2: *Medicare patients between 45 and 65 years.* In Figure A-2 we show the explanations for the visits where the model predicts the inclusion of ICD-9 code *428.0: Congestive heart failure.*

Interestingly, among the explanations, stands out the importance of the term *chf*, a medical acronym for congestive heart failure. The CAML model learns to associate a higher probability of using the ICD-9 code *428.0* when this term is present in the clinical notes, even though, looking closer at the notes of the misclassified patients, this term is many times associated to past episodes or family history of congestive heart failure, due to the age (45-65 years) of the patients. Also the term *lasix*, a first-line agent used to threat

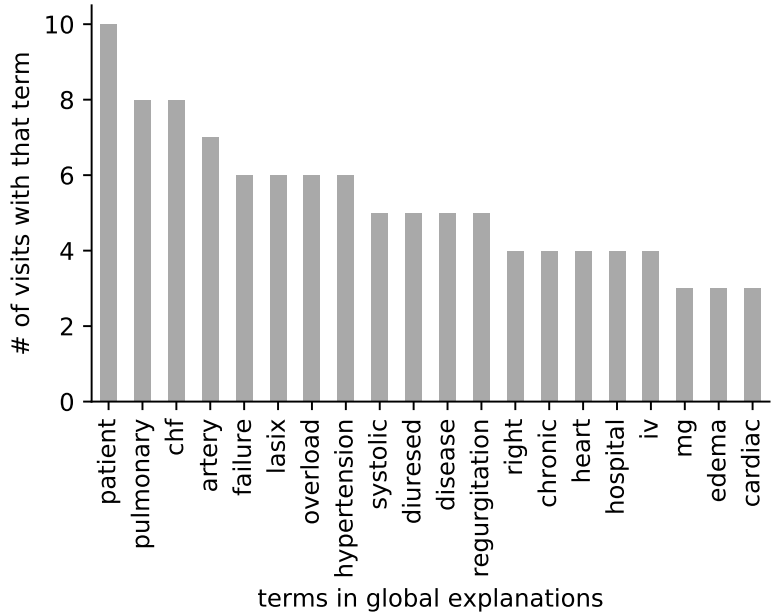


Figure A-2: Aggregated visualization of the relevant terms for the over-diagnosis of *428.0: Congestive heart failure* in Medicare patients between 45 and 65 years.

edema caused by congestive heart failure, has high importance for the model, even though the ICD-9 code *428.0* was not associated to many of these visits, but might be associated to past episodes (again due to the high age of the group).

B

Appendix B: user study additional information

B.1 Information sheet

You are invited to participate in a research study conducted by Andrea Beretta, Ph.D., from the National Research Council (CNR), Pisa, Italy, (andrea.beretta@isti.cnr.it) in collaboration with Cecilia Panigutti, Scuola Normale Superiore, Pisa, Italy, (cecilia.panigutti@sns.it)

Your participation is voluntary. You should read the information below and ask questions about anything you do not understand before deciding whether or not to participate in our study. You may also decide to discuss participation with your family or friends. You can download and print a copy of this form for your records.

B.1.1 What is the project's purpose?

The last decade has witnessed the rise of a black box society. Ubiquitous obscure algorithms, often based on sophisticated machine learning models trained on (big) data, which predict behavioral traits of individuals, such as credit risk, health status, personality profile. Many controversial cases have already highlighted that delegating decision-making to black-box algorithms is critical in many sensitive domains, including crime prediction, personality scoring, image classification, personal assistance, and more. The XAI project addresses the challenges of requiring that machine learning and AI be explainable and comprehensible in human terms. This is instrumental for validating the quality and correctness

of the resulting systems, and also for aligning the algorithms with human values and expectations, as well as preserving human autonomy and awareness in decision making. We are applying such technology in a variety of different domains. As part of this work, we need to understand if the produced outcome is comprehensible by professionals of a given domain and get their views on the usefulness of the provided explanations produced by the XAI tools. The project will run until 2024. The purpose of the study is to investigate user interaction with two different interfaces of an Artificial Intelligent system that provides you suggestion, and suggestion with explanation for advice in medical decisions. The results of this study may help us in our future work involving the design of explanations for Artificial Intelligent systems based on human and agent interaction.

B.1.2 Procedures

If you volunteer to participate in this study, we would ask you to do the following activities:

- **Pre-task survey following consent**: you will be asked to fill out a pre-task survey. This survey will give us information on your background, and demographic information
- **Interaction with two interfaces**: you will interact with an intelligent system that will provide you a suggestion, or a suggestion with an explanation. You will be asked to give two estimates according to your thoughts.
- **Post-game survey**: you will be asked to make ratings based on what you experienced during the task. Your overall commitment to this research is expected to be approximately 30 minutes.

Why have I been chosen?

You have been approached to participate due to your experience in the medical and healthcare field.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without any negative consequences. You do not have to

give a reason. If you wish to withdraw from the research, please contact Fosca Giannotti, ISTI-CNR, Via Moruzzi 1, 56124 Pisa, Italy. Email: fosca.giannotti@isti.cnr.it. Phone: +39 050621299.

What are the possible disadvantages and risks of taking part?

As your participation will be limited to very short periods and focused on your professional/citizen experience of understanding an explanation for a certain decision, no major disadvantages or risks are foreseen.

What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will result in the development of more accurate methods for automatic decisions. Many of these tools will be made available as open-source to the scientific community.

Will my taking part in this project be kept confidential?

All the information that we collect from you and about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications unless you have given your explicit consent for this on your participant consent form. If you agree to us sharing the information you provide with other researchers (e.g. by making it available in a data archive) then your personal details will not be included unless you explicitly request this.

What is the legal basis for processing my personal data?

According to data protection legislation, we are required to inform you that the legal basis we are applying in order to process your personal data is that 'processing is necessary for the performance of a task carried out in the public interest' (Article 6(1)(e)).

What will happen to the data collected, and the results of the research project?

We wish to inform You that, according to the regulation in force the processing of Your personal data will be based on the principles of correctness, lawfulness, and transparency as well as the protection of confidentiality and Your rights. Therefore we provide YOU with the following information:

B.1.3 Study information

- Data Controller: Italian National Research Council (CNR), (Piazzale Aldo Moro, 7 - 00185 Roma, Italy), through Institute of Information Science and Technologies "A. Faedo" (ISTI).
- Data processor: Director of ISTI CNR in Pisa, via Moruzzi 1, 56124, Pisa, e-mail address: direttore@isti.cnr.it
- Data Protection Officer of CNR: e-mail address: rpd@cnr.it
- Project Manager: e-mail address: fosca.giannotti@isti.cnr.it

Your data will be processed through electronic, automated and/or manual instruments, with methods and tools to ensure maximum security and confidentiality, by authorized personnel in compliance with the regulations in force and following the operating instructions provided for by the regulations of the structure.

Your personal data being processed and the other information acquired will be stored and processed for the sole purpose of the project.

Due to the nature of this research, it is very likely that other researchers may find the data collected to be useful in answering future research questions. We will ask for your explicit consent for your data to be shared in this way.

However, they may be used only for scientific research purposes even after the end of the project in compliance with the "Code of conduct and good conduct for the processing of scientific and statistical data".

Your personal data being processed and other information acquired will be retained

- in database and servers of CNR-ISTI;

- using cloud services provided by third parties, under contractual agreements for the protection of personal data;
- in database and servers of the partners of the project: Department of Compiuter Science, Unversity of Pisa, under contractual agreements for the protection of personal data;

Who is organising and funding the research?

The XAI project has received funding from the European Union/EU under the Information and Communication Technologies (ICT) theme of the Horizon 2020 Programme for R&D, grant XAI (825297).

Who has ethically reviewed the project?

This project has been ethically approved via the ethical commission of CNR
https://www.cnr.it/it/ethical-clearance
cnr.ethics@cnr.it

What if something goes wrong and I wish to complain about the research?

Any complaints by participants will be handled. In the first instance, you may contact the Principal Investigator, Fosca Giannotti (fosca.giannotti@isti.cnr.it). In case you feel the complaint is not been handled to your satisfaction (e.g. by the Principal Investigator) you can contact the director of ISTI, Roberto Scopigno, who will then escalate the complaint through the appropriate channels.

Contact for further information

Fosca Giannotti, ISTI-CNR, Via Moruzzi 1, 56124 Pisa, Italy. Email: fosca.giannotti@isti.cnr.it. Phone: +39 050621299.

You will be given a copy of the information sheet and a signed consent form to keep.

Thank you for taking part in the project.

B.2 Informed consent

Dear Sir/Madam,

we wish to inform You that, according to the regulation in force the processing of Your personal data will be based on the principles of correctness, lawfulness and transparency as well as the protection of confidentiality and Your rights. Therefore, we provide YOU with the following information:

- Data Controller: Italian National Research Council (CNR), (Piazzale Aldo Moro, 7 - 00185 Roma, Italy), e-mail address: presidente@cnr.it
- Data processor: Director of Institute of Information Science and Technologies "A. Faedo" (ISTI).CNR in Pisa, via Moruzzi 1, 56124, Pisa, e-mail address: direttore@isti.cnr.it delegated by president of CNR
- Data Protection Officer of CNR: Dr. Raffaele Conte, e-mail address: rpd@cnr.it
- Project Manager and contact person: e-mail address: fosca.giannotti@isti.cnr.it authorized by Director of ISTI

PURPOSES AND THE LEGAL BASIS FOR THE PROCESSING

XAI: Science and technology for the eXplanation of AI decision making, GAP-834756, is a 60-months EU-funded project addressing the challenges of requiring that decisions suggested by autonomous intelligent systems be comprehensible in human terms. The decision logic of modern decision support systems is often based on sophisticated models inferred from large data sets of examples (Big data); the problem lies in the fact that the rationale of the suggested choice remains obscure and unintelligible in human terms such as for instance a system that suggests denying a mortgage application without explaining the reasons. Many controversial cases have already highlighted that delegating decision-making to "black-box algorithms" is critical in many sensitive domains, including crime prediction, personality scoring, image classification, personal assistance, and more. The XAI project has developed methods that address the challenges of requiring that decision support systems based on Artificial Intelligence be explainable and comprehensible in human terms. This is fundamental for aligning the algorithms with human values and expectations, as well as preserving human autonomy and awareness in decision making.

This survey is aimed at validating the quality of the explanation the technology XAI has produced.

METHODS OF DATA PROCESSING

The processing will be carried out by electronic, automated, and/or manual instruments, with methods and tools to ensure maximum security and confidentiality, by authorized personnel in compliance with the regulations in force and following the operating instructions provided for by the regulations of the structure.

DATA RETENTION

Your personal data subject to processing and other information acquired will be stored and processed for the sole purpose and duration of the project and successive 5 years (project starts in September 2019 and ends August 2024). However, they may be used only for scientific research purposes even after the end of the project in compliance with the "Code of conduct and good conduct for the processing of scientific and statistical data" coherently General Data 14 Protection Regulation (EU) 2016/679 "GDPR" and other relevant national laws and regulations. Your personal data being processed and other information acquired will be retained in the database and servers of CNR-ISTI.

COMMUNICATION AND DATA DISSEMINATION

Your personal data will not be disclosed to other subjects. Your personal data will not be disseminated. ISTI – CNR, in accordance with the "Code of conduct and professional practice applying to the processing of personal data for statistical and scientific purposes (Published in the Official Journal no. 190 of August 14, 2004)" Garante della Privacy, has, however, the possibility to share with the scientific community, for research purposes, aggregates, statistics, results of the analysis. These results will be used anonymously for scientific dissemination.

TRANSFER OF PERSONAL DATA

Data will not be transferred to non-EU countries. In any case, it ensured from now on that the transfer will take place in accordance with the applicable legal provisions and the standard contractual clauses provided by the European Commission, in order to guarantee

compliance with the principles of lawfulness and adequacy of treatment as provided for by the EU GDPR 2016/679 on the base of the safe list of the countries provided by the EU commission.

RIGHT OF ACCESS BY THE DATA SUBJECT

Pursuant to art. 15 of the EU Reg., You have the right to access the data being processed, including the right to receive a copy. These include the expected retention period or, if this is not possible, the criteria used to define this period, as well as the guarantees applied in case of transfer of data to third countries. Where applicable, You also have the rights referred to in Articles 16-21 of the GDPR. 2016/679 (Right of rectification, right to be forgotten, right of limitation of treatment, right to data portability, right of opposition), as well as the right to lodge a complaint with a supervisory authority.

CONSENT TO THE PROCESSING OF PERSONAL DATA FOR THE IMPLEMENTATION OF PROJECT ACTIVITIES

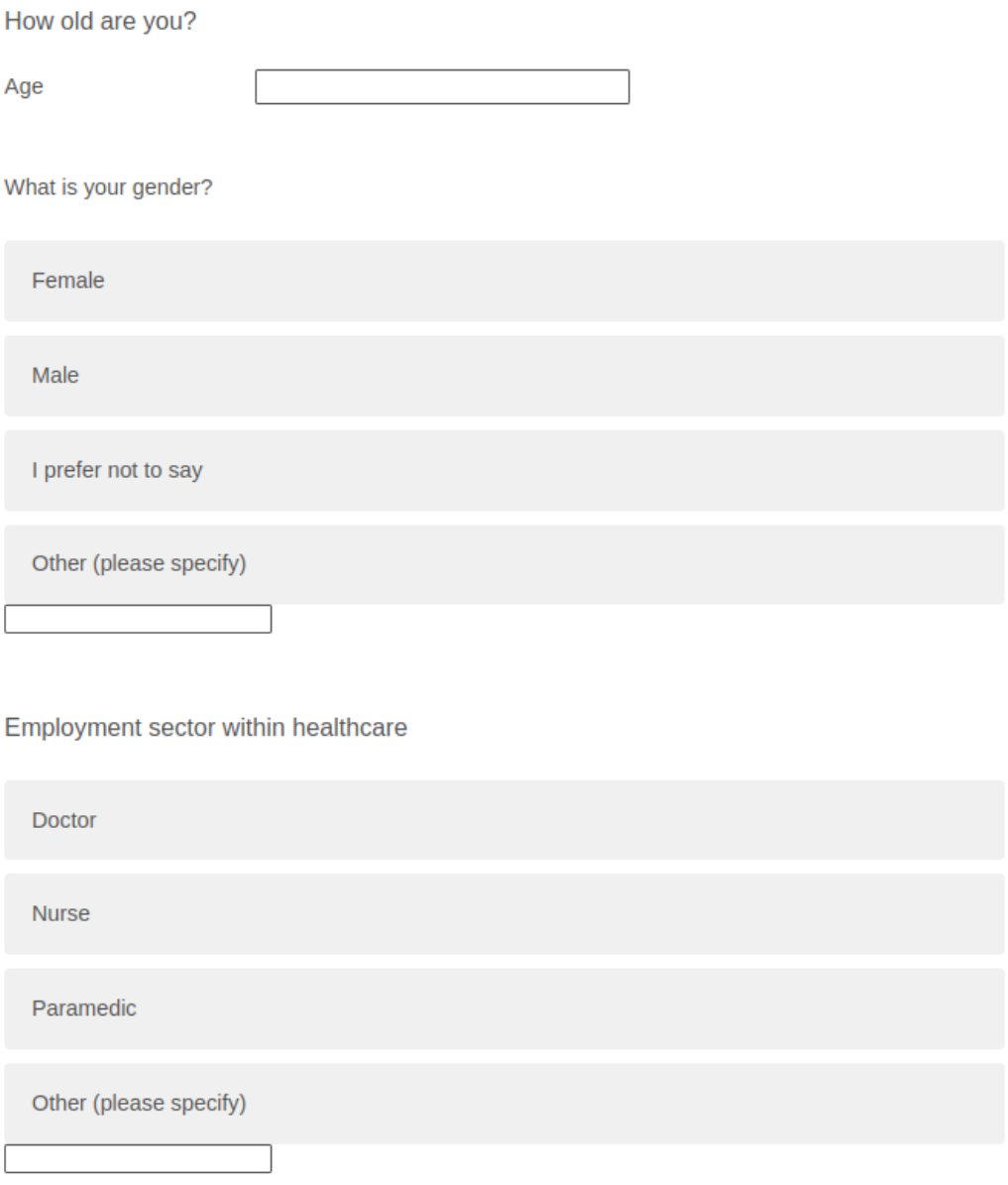
The provision of data is optional; however, any refusal to provide it, or to subsequently deny the processing of data already provided, could totally or partially compromise the outcome of the project.

The undersigned declares to have read the above information and to consent to the processing of their data necessary for the realization of the purposes envisaged by the project.

- I do accept
- I do not accept

B.3 Demographics

This part is dedicated to collecting your demographic data to cluster the responses of each participant in the research. Demographics will be important for the following survey. Please be aware that to prevent random responses, we used attention checks. Thank you for your participation. Please answer each question as accurately as possible by selecting the correct answer or filling in the space provided.



B.4 Need for cognition (NFC)

For each of the statements below, please indicate whether or not the statement is characteristic of you or of what you believe. For example, if the statement is extremely uncharacteristic of you or of what you believe about yourself (not at all like you) please select "Strongly disagree" on the line to the left of the statement. If the statement is extremely characteristic of you or of what you believe about yourself (very much like you) please select "Strongly agree" on the line to the left of the statement. You should use the following scale as you rate each of the statements below.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I would prefer complex to simple problems.	○	○	○	○	○
I like to have the responsibility of handling a situation that requires a lot of thinking.	○	○	○	○	○
Thinking is not my idea of fun.	○	○	○	○	○
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.	○	○	○	○	○
I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.	○	○	○	○	○
I find satisfaction in deliberating hard and for long hours.	○	○	○	○	○
I only think as hard as I have to.	○	○	○	○	○
I prefer to think about small, daily projects to long-term ones.	○	○	○	○	○
I like tasks that require little thought once I've learned them.	○	○	○	○	○
The idea of relying on thought to make my way to the top appeals to me.	○	○	○	○	○
I really enjoy a task that involves coming up with new solutions to problems.	○	○	○	○	○
Learning new ways to think doesn't excite me very much.	○	○	○	○	○
I prefer my life to be filled with puzzles that I must solve.	○	○	○	○	○

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The notion of thinking abstractly is appealing to me.	○	○	○	○	○
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	○	○	○	○	○
I feel relief rather than satisfaction after completing a task that required a lot of mental effort.	○	○	○	○	○
It's enough for me that something gets the job done; I don't care how or why it works.	○	○	○	○	○
I usually end up deliberating about issues even when they do not affect me personally.	○	○	○	○	○
Click on the lowest grade of the scale, this is to prevent random clicking	○	○	○	○	○

B.5 Instructions

The goal of our research is to **compare two interfaces of an Artificial Intelligence (AI)** system for **clinical decision support**.

You will see two types of the interface of the **same AI algorithm**:

- **Interface A**: the AI algorithm provides a suggestion for a patient case
- **Interface B**: the AI algorithm provides a suggestion for a patient case AND an explanation of its internal decision-making process

---

To evaluate which interface you prefer and which one is more valuable for clinical decision support, we will ask you to perform an **estimation task** using the **two interfaces**. Then we will ask you to answer some questions.

The estimation task consists of estimating the **chances that a patient will have an Acute Myocardial Infarction (Acute MI) in the near future** based on their past clinical history.

For **each patient and each interface**, you will be asked to estimate their chances of developing an Acute MI **twice**:

- The first time you will estimate their chances **based only on the clinical history** of the patient and your **knowledge and experience**.
- The second time you will receive a **suggestion based on an AI algorithm** and you can decide whether to follow the suggestion and change your initial estimate or not.

Before we begin, please tell us: are you familiar with estimating the chances that a patient will suffer from an Acute Myocardial Infarction (Acute MI)?

Not at all familiar	Slightly familiar	Somewhat familiar	Moderately familiar	Extremely familiar

This study will ask you to estimate patients' chances of suffering from an Acute MI in the near future. Just do your best.

In the real world, this is a very important estimate

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

In the real world, if I make the wrong estimate, I will lose a lot.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

Throughout this survey, we will ask you to write your estimate in a form like the one below. Try it out writing "50" to indicate a 50% chance of developing an Acute MI.

Chance of suffering from an acute myocardial infarction in the future (%)

We will ask you to use a sliding scale like the ones below to indicate your confidence level about your estimate. Try out the examples below.

Set this slider to 100.

0 10 20 30 40 50 60 70 80 90 100

Set this slider to 100.

Set this slider to 0

0 10 20 30 40 50 60 70 80 90 100

Set this slider to 0.

B.6 Tutorial clinical history

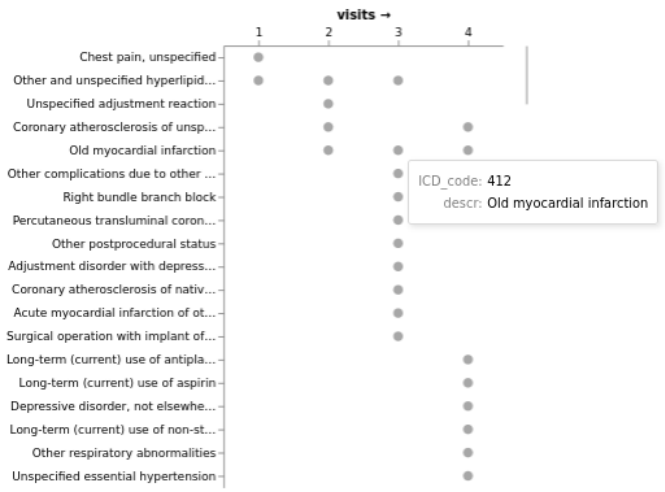
Now you will be presented with a brief tutorial to familiarize yourself with the representation of **patients' clinical history** used in the estimation task. You will be asked some questions to make sure you understand such representation.

You will be asked to make your **initial estimate** based on a **patient clinical history** like the one portrayed below.

This patient has **4 visits** in their clinical history. In each visit, they were diagnosed with one or more conditions.

You can explore the description of each condition by moving the cursor over the **grey dots representing the diagnoses of each visit**. For example, this patient was diagnosed with 8 conditions (represented by the grey dots) in their 4th visit. **The order of the conditions in each visit is not important**.

Move your mouse cursor over the second dot from the top. You can see the ICD9 code of the condition (International Classification of Diseases, 9th revision) **412** and its description **"Old myocardial infarction."**



What are the codes of the conditions diagnosed in the first visit of the patient?

272.4 and 309.0

272.4 and 786.50

272.4 and 412 and 414

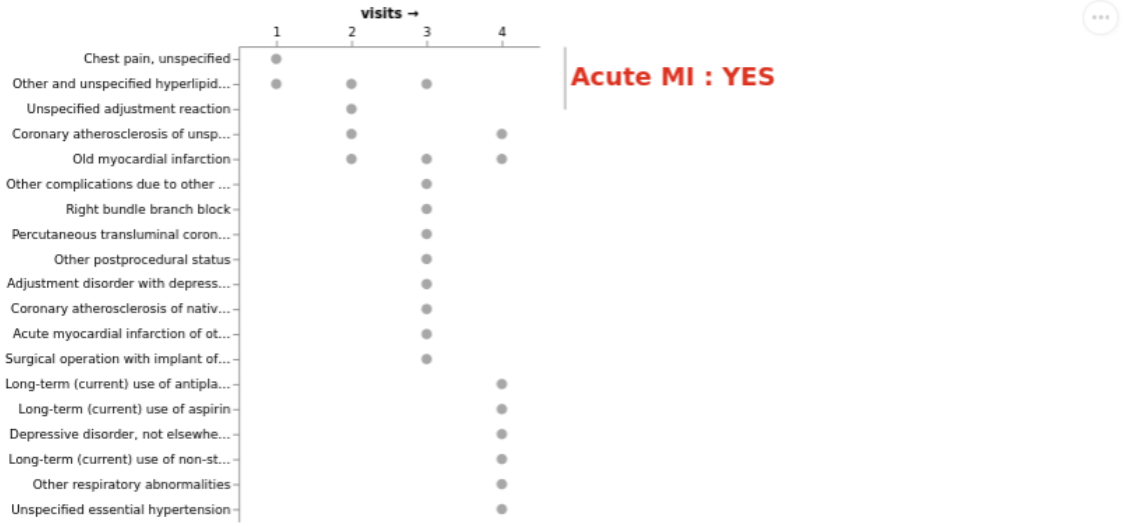
B.7 Tutorial Interface A: only suggestion of the AI algorithm

Now you will be presented with a brief tutorial to familiarize yourself with **interface A: only suggestion of the AI algorithm**.

Please read the instructions carefully.

This interface shows you a **suggestion**. This suggestion is made by an algorithm that has learned to predict if a patient will have an **Acute Myocardial Infarction (Acute MI)** in the near future based on their clinical history.

For example, for this patient, the algorithm predicts that this patient will have an Acute MI in the near future: "**Acute MI: YES**": you might need to keep them overnight for observation at the hospital.

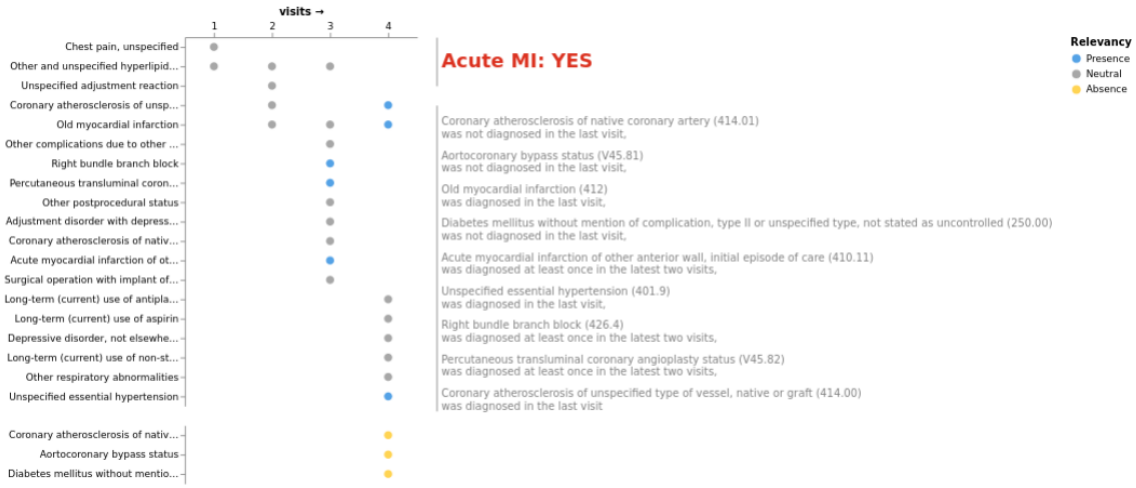


B.8 Tutorial Interface B: suggestion of the AI algorithm and explanation

Now you will be presented with a brief tutorial to familiarize yourself with **interface B: suggestion of the AI algorithm and explanation** of its internal decision-making process.

You will be asked some questions to make sure you understand the explanation.
Please read the instructions carefully.

This interface shows you a **suggestion and an explanation for such a suggestion**. This suggestion is made by an algorithm that has learned to predict if a patient will have an **Acute Myocardial Infarction (Acute MI)** in the near future based on their clinical history.

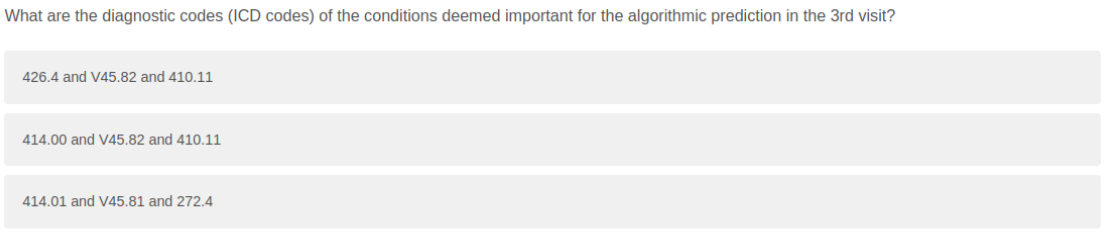


For example, for this patient, the algorithm predicts that this patient will have an Acute MI in the near future: **"Acute MI: YES"**. You might need to keep them overnight for observation at the hospital.

The **explanation** for such a prediction is written under the suggestion. You can drag the cursor over each part of the explanation to highlights the history of the conditions deemed important by the algorithm.

The algorithm considered important the fact that the patient **had some conditions diagnosed** in the past (**blue dots**) AND the fact the patient **did not have some other conditions** in their clinical history (**yellow dots**). Other conditions were deemed **not relevant** for the algorithmic prediction (**grey dots**).

For example, in this case, the algorithm deemed it important that the patient was diagnosed with "Unspecified essential hypertension" in the last visit (coded as 401.9 in the 9th revision of the International Classification of Diseases). You can also see that the algorithm deemed important the fact that the patient was NOT diagnosed with a "Coronary atherosclerosis of native coronary artery" in the last visit.



B.9 Estimation task

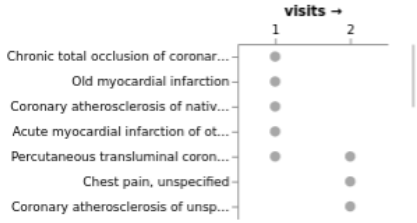
You will now be presented with the **real estimation task**.

You will be asked to give your **first estimate** based only on the clinical history of the patient. Then you will be asked to provide a **final estimate** after the suggestion given by the AI algorithm (interface A/B).

B.9.1 Initial estimate

Considering this patient with a 2-visits clinical history, please indicate their chances of having an Acute MI in the near future.

Please use your mouse cursor to explore the diagnoses of the clinical history to inform your estimate.



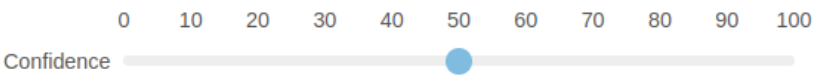
The patient has a chance between 0 and 100 of suffering from an Acute MI in the near future.

Make your best **initial estimate** between 0 and 100:

Chance of suffering from an acute myocardial infarction in the future (%)

Your **initial estimate** is %.

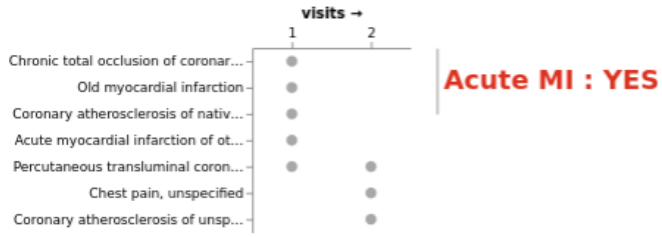
How confident are you in your estimate? Please indicate your confidence level using the slider below (0 = not confident at all, 50 = somewhat confident, 100 = completely confident).



B.9.2 Final estimate: only suggestion

The algorithm predicts that **this patient will have an Acute MI in the near future**: you might need to keep them overnight for observation at the hospital.

The algorithm uses the patient's clinical history to make its prediction. You can choose to follow its suggestion or not by changing your initial estimate.



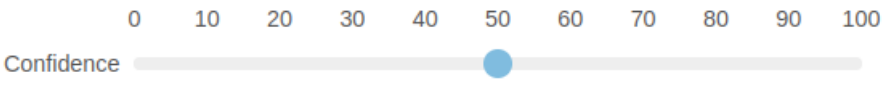
The patient has a chance between 0 and 100 of suffering from an Acute MI in the near future.

Make your best **final estimate** between 0 and 100:

Chance of suffering from an acute myocardial infarction in the future (%)

Your **final estimate** is %.

How confident are you in your estimate? Please indicate your confidence level using the slider below
(0 = not confident at all, 50 = somewhat confident, 100 = completely confident).



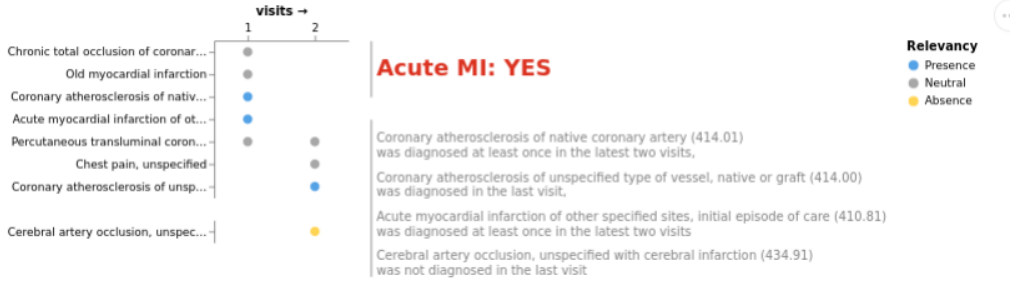
B.9.3 Final estimate: suggestion and explanation

The algorithm predicts that **this patient will have an Acute MI in the near future**: you might need to keep them overnight for observation at the hospital. The algorithm uses the patient's clinical history to make its prediction.

You can see the conditions that were deemed important by the algorithm highlighted on the clinical history of the patient (blue = it was important that this condition was present in the patient's clinical history, yellow = it was important that this condition was NOT present in the patient's clinical history).

You can also see a summary of the explanation written under the algorithmic prediction.

You can choose to follow its suggestion or not by changing your initial estimate.



The patient has a chance between 0 and 100 of suffering from an Acute MI in the near future.

Make your best **final estimate** between 0 and 100:

Chance of suffering from an acute myocardial infarction in the future (%)

Your **final estimate** is %.

How confident are you in your estimate? Please indicate your confidence level using the slider below
(0 = not confident at all, 50 = somewhat confident, 100 = completely confident).

B.10 Trust scale

We ask you to answer a short questionnaire to investigate your perceptions regarding the interface you just used.

	I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly
I am **confident** in the algorithm. I feel it works well.	○	○	○	○	○
The outputs of the algorithm are very **predictable**	○	○	○	○	○
The tool is very **reliable**. I can count on it to be correct all the time.	○	○	○	○	○
I feel safe that when I rely on the algorithm I will get the **right** answers	○	○	○	○	○
The algorithm is efficient in that it works very **quickly**	○	○	○	○	○
I am **wary** of the algorithm	○	○	○	○	○
The app can perform the task **better than a novice** human user.	○	○	○	○	○
I **like using** the system for decision making	○	○	○	○	○

B.11 UTAUT - pt1

We ask you to answer a short questionnaire to investigate your perceptions regarding the interface you just used.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I would find the system **useful** in my job	○	○	○	○	○
Using the system would enable me to **accomplish** tasks **more quickly**	○	○	○	○	○
Using the system **would increase my productivity**	○	○	○	○	○
If I use the system, I would increase my chances of **getting a raise**	○	○	○	○	○
My interaction with the system is **clear and understandable**	○	○	○	○	○
Interacting with the system does **not** require a lot of my **mental effort**	○	○	○	○	○
I find the system to be **easy to use**	○	○	○	○	○
I find it **easy** to get the system to do what I want to do	○	○	○	○	○
Using the system is a **good idea**	○	○	○	○	○
The system would make work **more interesting**	○	○	○	○	○
Working with the system would be **fun**	○	○	○	○	○
I **would like working** with the system	○	○	○	○	○

B.12 UTAUT - pt2

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
People **who influence** my behavior would think that **I should use the system**	○	○	○	○	○
People **who are important** to me would think that **I should use the system**	○	○	○	○	○
The senior **management** of this business **would be helpful in the use of the system**	○	○	○	○	○
In general, **the organization would support the use** of the system	○	○	○	○	○
I have **the resources necessary** to use the system	○	○	○	○	○
I have **the knowledge necessary** to use the system	○	○	○	○	○
The system is **not compatible** with other systems I use	○	○	○	○	○
A specific person (or group) would be **available for assistance** with system difficulties	○	○	○	○	○
Doctors who use this system would have **more prestige** than those who do not	○	○	○	○	○
Doctors who use this system **would have a high profile**	○	○	○	○	○
Having this system will be a **status symbol**	○	○	○	○	○
In my job, usage of the system **would be important**	○	○	○	○	○

B.13 UTAUT - pt3

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
In my job, usage of the system **would be relevant**	○	○	○	○	○
The use of the system **is pertinent** to my various job-related tasks	○	○	○	○	○
The **quality of the output** I get from the system is high	○	○	○	○	○
I have **no problem with the quality** of the system's output	○	○	○	○	○
I rate the results from the system to be **excellent**	○	○	○	○	○
I have **no difficulty telling others about the results** of using the system	○	○	○	○	○
I believe I could communicate to others **the consequences** of using the system	○	○	○	○	○
The results of using the system are **apparent** to me	○	○	○	○	○
I would **have difficulty explaining** why using the system may or may not be beneficial	○	○	○	○	○
I **intend to use** the system if the system would be accessible to me	○	○	○	○	○
I **predict** I would use the system if the system would be accessible to me	○	○	○	○	○
I **plan** to use the system if the system would be accessible to me	○	○	○	○	○

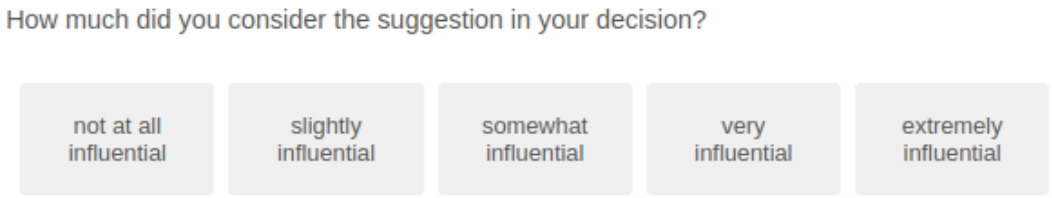
B.14 Explanation satisfaction scale

We ask you to answer a short questionnaire to investigate your perceptions regarding the explanation of the algorithmic prediction given by the system you just used

	I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly
From the explanation, I **understand** how the algorithm works.	○	○	○	○	○
This explanation of how the algorithm works is **satisfying**.	○	○	○	○	○
This explanation of how the algorithm works has **sufficient detail**.	○	○	○	○	○
This explanation of how the algorithm works seems **complete**.	○	○	○	○	○
This explanation of how the algorithm works **tells me how to use it**.	○	○	○	○	○
This explanation of how the algorithm works is **useful to my goals**.	○	○	○	○	○
This explanation of the algorithm shows me how **accurate** the algorithm is.	○	○	○	○	○
This explanation lets me judge when I should **trust and not trust** the algorithm	○	○	○	○	○

B.15 Open questions

- What was your overall impression of the AI interface you just used?
- What was the thing you prefer the most about this AI interface?
- What was the thing you dislike the most about this AI interface?
- Have you found any difficulties? If yes, specify what they were
- How would you change this AI interface?
- How do you think the AI system works (builds its suggestion)?



B.15.1 Additional open questions for interface B "Suggestion and explanation"

- In your own words, could describe what an explanation is?
- According to your thoughts, please describe what features the explanation is based on.
- What can you suggest to improve the explanation?

B.16 Final open questions

- How did the suggestion provided by interface A "only suggestion" influenced your final estimate?
- How did the suggestion provided by interface B "suggestion and explanation" influenced your final estimate?
- (Optional) Do you have any feedback about the experiment?

Bibliography

[1] Charter of fundamental rights of the european union.

[2] Eu general data protection regulation.

[3] Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices.

[4] Stakeholder consultation on guidelines' first draft.

[5] The eu wants to become the world's super-regulator in ai. *The Economist*, 2021.

[6] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of training diagnostic algorithms on data with demographic bias, 2020.

[7] Abdelhakeem MB Abdelrahman, Ahmad Kayed, et al. A survey on semantic similarity measures between concepts in health domain. *American Journal of Computational Mathematics*, 5(02):204, 2015.

[8] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 42–46, 2017.

[9] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):1–8, 2018.

[10] Barbara D Adams, Lora E Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann. Trust in automated systems. *Ministry of National Defence*, 2003.

[11] Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

[12] Julius A Adebayo et al. *FairML: ToolBox for diagnosing bias in predictive modeling.* PhD thesis, Massachusetts Institute of Technology, 2016.

[13] Bibb Allen, Sheela Agarwal, Laura Coombs, Christoph Wald, and Keith Dreyer. 2020 acr data science institute artificial intelligence survey. *Journal of the American College of Radiology*, 2021.

[14] Ahmad Fayez S Althobaiti. Comparison of ontology-based semantic-similarity measures in the biomedical text. *Journal of Computer and Communications*, 5(02):17, 2017.

[15] Elvio Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479, 2021.

[16] Robert Andrews, Joachim Diederich, and Alan B Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, 8(6):373–389, 1995.

[17] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1):8753–8830, 2017.

[18] Sule Anjomshoae, Timotheus Kampik, and Kary Främling. Py-ciu: A python library for explaining machine learning predictions using contextual importance and utility. In *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*, 2020.

[19] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[20] Alice Armitage, Andrew Cordova, and Rebecca Siegel. Design-thinking: The answer to the impasse between innovation and regulation. *UC Hastings Research Paper*, (250), 2017.

[21] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv preprint arXiv:2008.02766*, 2020.

[22] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[23] Robert Avram, Jeffrey E Olgin, Peter Kuhar, J Weston Hughes, Gregory M Marcus, Mark J Pletcher, Kirstin Aschbacher, and Geoffrey H Tison. A digital biomarker of diabetes from smartphone-based vascular signals. *Nature Medicine*, 26(10):1576–1582, 2020.

[24] Babylon Health. Babylon health.

[25] Ji-Won Baek, Joo-Chang Kim, Junchul Chun, and Kyungyong Chung. Hybrid clustering based health decision-making for improving dietary habits. *Technology and Health Care*, (Preprint):1–14, 2019.

[26] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[27] Tian Bai et al. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD*, pages 43–51. ACM, 2018.

[28] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, pages 1–10, 2021.

[29] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[30] Andrea Barucci and Emanuele Neri. Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning, 2020.

[31] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[32] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

[33] Andrew L Beam, Arjun K Manrai, and Marzyeh Ghassemi. Challenges to the reproducibility of machine learning models in health care. *JAMA*, 2020.

[34] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[35] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[36] Valentina Bellemo, Zhan W Lim, Gilbert Lim, Quang D Nguyen, Yuchen Xie, Michelle YT Yip, Haslina Hamzah, Jinyi Ho, Xin Q Lee, Wynne Hsu, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in africa: a clinical validation study. *The Lancet Digital Health*, 1(1):e35–e44, 2019.

[37] João Bento, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. *arXiv preprint arXiv:2012.00073*, 2020.

[38] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370. AAAI Press, 1994.

[39] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

[40] Alan F. Blackwell. Ethnographic artificial intelligence. *Interdisciplinary Science Reviews*, 46(1-2):198–211, 2021.

[41] Natalia Blanco, Lyndsay M O'Hara, Gwen L Robinson, Jeanine Brown, Emily Heil, Clayton H Brown, Brian D Stump, Bryant W Sigler, Anusha Belani, Heidi L Miller, et al. Health care worker perceptions toward computerized clinical decision support tools for clostridium difficile infection reduction: a qualitative study at 2 hospitals. *American journal of infection control*, 46(10):1160–1166, 2018.

[42] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[43] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*, 2021.

[44] Francesco Bodria, A. Panisson, A. Perotti, and Simone Piaggesi. Explainability methods for natural language processing: Applications to sentiment analysis. In *Proceedings of the 28th Italian Symposium on Advanced Database Systems (SEBD 2020)*, 2020.

[45] Adam S Bodzin and Talia B Baker. Liver transplantation today: where we are now and where we are going. *Liver Transplantation*, 24(10):1470–1475, 2018.

[46] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[47] Anu Bradford. The brussels effect. *Nw. UL Rev.*, 107:1, 2012.

[48] Andrea Brennen. What do people really want when they say they want" explainable ai?" we asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2020.

[49] M Brounillette. Deep learning is a black box, but health care won't mind, 2017.

[50] Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, and Roger Wattenhofer. On the validity of self-attention as explanation in transformer models. *arXiv preprint arXiv:1908.04211*, 2019.

[51] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[52] Jacquelyn Burkell and Alexandre Fortier. Could we do better? behavioural tracking on recommended consumer health websites. *Health Information & Libraries Journal*, 32(3):182–194, 2015.

[53] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE, 2015.

[54] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. The efficient assessment of need for cognition. *Journal of personality assessment*, 48(3):306–307, 1984.

[55] Béatrice Cahour and Jean-François Forzy. Does projection into use improve trust and exploration? an example with a cruise control system. *Safety science*, 47(9):1260–1270, 2009.

[56] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[57] David Capper, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, et al. DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.

[58] Donna J Cartwright. Icd-9-cm to icd-10-cm codes: what? why? how?, 2013.

[59] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[60] Joan A Casey, Brian S Schwartz, Walter F Stewart, and Nancy E Adler. Using electronic health records for population health research: a review of methods and applications. *Annual review of public health*, 37:61–81, 2016.

[61] Ross Casey. IBM's watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show. 2018.

[62] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[63] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

[64] Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi. Treating health disparities with artificial intelligence. *Nature Medicine*, 26(1):16–17, 2020.

[65] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*, 2020.

[66] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[67] Yongxi Chen and Anne SY Cheung. The transparent self under big data profiling: Privacy and chinese legislation on the social credit system. 2017.

[68] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2019.

[69] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, 392(10162):2388–2396, 2018.

[70] Benjamin Chin-Yee and Ross Upshur. The impact of artificial intelligence on clinical judgment: A briefing document. 2020.

[71] chinainnovationfunding.eu. State council's plan for the development of new generation artificial intelligence.

[72] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[73] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.

[74] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[75] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Multi-label transfer learning for the early diagnosis of breast cancer. *Neurocomputing*, 392:168–180, 2020.

[76] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[77] Amanda Clare and Ross D King. Knowledge discovery in multi-label phenotype data. In *European conference on principles of data mining and knowledge discovery*, pages 42–53. Springer, 2001.

[78] I Glenn Cohen and Michelle M Mello. Hipaa and protecting health information in the 21st century. *Jama*, 320(3):231–232, 2018.

[79] Giovanni Comandé. Multilayered (accountable) liability for artificial intelligence. In Dirk Staudenmayer Sebastian Lohsse, Reiner Schulze, editor, *Liability for Artificial Intelligence and the Internet of Thing*, pages 165–187, 2019.

[80] Giovanni Comandé and Gianclaudio Malgieri. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, 7(4):243–265, 2017.

[81] European Commission, Directorate-General for Research, and Innovation. *Gender intersectional bias in artificial intelligence*. Publications Office, 2020.

[82] Roberto Confalonieri, Tillman Weyde, Tarek R Besold, and Fermín Moscoso del Prado Martín. Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks. 2020.

[83] Oscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. Methodologies, tools and languages for building ontologies. where is their meeting point? *Data & knowledge engineering*, 46(1):41–64, 2003.

[84] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[85] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878*, 2020.

[86] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30, 1995.

[87] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, pages 24–30, 1996.

[88] Rogier Creemers. Planning outline for the construction of a social credit system (2014-2020). *China Copyright and Media*, 14, 2014.

[89] Valerie Cross and Youbo Wang. Semantic relatedness measures in ontologies using information content and fuzzy set theory. In *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.*, pages 114–119. IEEE, 2005.

[90] Christine M Cutillo, Karlie R Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, and Kenneth D Mandl. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine*, 3(1):1–5, 2020.

[91] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[92] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[93] Francesco De Comité, Rémi Gilleron, and Marc Tommasi. Learning multi-label alternating decision trees from texts and data. In *International workshop on machine learning and data mining in pattern recognition*, pages 35–49. Springer, 2003.

[94] Jean Decety. Empathy in medicine: what it is, and how much we really need it. *The American journal of medicine*, 133(5):561–566, 2020.

[95] Berkeley J Dietvorst and Soaham Bharti. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science*, 31(10):1302–1314, 2020.

[96] Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728, 2021.

[97] Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

[98] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[99] Jinyun Duan, Yue Xu, and Lyn M Van Swol. Influence of self-concept clarity on advice seeking and utilisation. *Asian Journal of Social Psychology*, 2020.

[100] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[101] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. Fairecsys: Mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, 9(2):197–213, 2020.

[102] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[103] Fabian Eitel, Kerstin Ritter, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2019.

[104] Shaker El-Sappagh and Farman Ali. Ddo: a diabetes mellitus diagnosis ontology. In *Applied Informatics*, volume 3, page 5. Springer, 2016.

[105] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2002.

[106] Anne Elixhauser. *Clinical Classifications for Health Policy Research: Version 2: Software and User's Guide*. Number 96. US Department of Health and Human Services, Public Health Service, Agency . . . , 1996.

[107] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[108] MedTech Europe. Medtech europe responds to the pilot on the trustworthy artificial intelligence assessment list.

[109] European Commission. CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION.

[110] European Commission. Artificial Intelligence for Europe, 2018.

[111] European Commission. White paper: On Artificial Intelligence - A European approach to excellence and trust, 2020.

[112] European Parliament. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 04 2021.

[113] Lesley J Fallowfield. Treatment decision-making in breast cancer: the patient–doctor relationship. *Breast cancer research and treatment*, 112(1):5–13, 2008.

[114] Wenjuan Fan, Jingnan Liu, Shuwan Zhu, and Panos M Pardalos. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (aimdss). *Annals of Operations Research*, pages 1–26, 2018.

[115] FDA. Proposed Regulatory Framework for Modification to Artificial Intelligence/-Machine Learning (AI/ML)-based Software as a Medical Device.

[116] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[117] Ruiwei Feng, Yan Cao, Xuechen Liu, Tingting Chen, Jintai Chen, Danny Z Chen, Honghao Gao, and Jian Wu. Chronet: A multi-task learning based approach for prediction of multiple chronic diseases. *Multimedia Tools and Applications*, pages 1–15, 2021.

[118] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.

[119] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, pages 283–286, 2020.

[120] Luciano Floridi. On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, 29(4):307–312, 2016.

[121] Brian J Fogg and Hsiang Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87, 1999.

[122] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[123] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Advances in neural information processing systems*, pages 2053–2061, 2015.

[124] Sidney Fussell. Google's totally creepy, totally legal health-data harvesting. *The Atlantic*, 2019.

[125] Jingyue Gao, Xiting Wang, Yasha Wang, Zhao Yang, Junyi Gao, Jiangtao Wang, Wen Tang, and Xing Xie. Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1036–1041. IEEE, 2019.

[126] David García-Soriano and Francesco Bonchi. Fair-by-design matching. *Data Mining and Knowledge Discovery*, pages 1–45, 2020.

[127] Ruiquan Ge, Renfeng Zhang, and Pu Wang. Prediction of chronic diseases with multi-label neural network. *IEEE Access*, 8:138210–138216, 2020.

[128] Michael Geruso and Timothy Layton. Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.

[129] Fosca Giannotti, Dino Pedreschi, and Cecilia Panigutti. Ia comprensibile per il supporto alle decisioni: Doctor xai. In *Biopolitica, pandemia e democrazia. Rule of law nella società digitale. Vol. 3: Pandemia e tecnologie. L'impatto su processi, scuola e medicina*, pages 109–120. Il mulino, 2021.

[130] Francesca Gino and Maurice E Schweitzer. Take this advice and shove it. In *Academy of Management Proceedings*, volume 2008, pages 1–5. Academy of Management Briarcliff Manor, NY 10510, 2008.

[131] Dominic Girardi, Sandra Wartner, Gerhard Halmerbauer, Margit Ehrenmüller, Hilda Kosorus, and Stephan Dreiseitl. Using concept hierarchies to improve calculation of patient similarity. *Journal of biomedical informatics*, 63, 2016.

[132] Fausto Giunchiglia and Ilya Zaihrayeu. *Lightweight Ontologies*, pages 1613–1619. Springer US, Boston, MA, 2009.

[133] Tasha Glenn and Scott Monteith. Privacy in the digital world: medical and health data outside of hipaa protections. *Current psychiatry reports*, 16(11):494, 2014.

[134] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

[135] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[136] Google. Artificial Intelligence at Google: Our Principles.

[137] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[138] DM Gruen, S Chari, MA Foreman, O Seneviratne, R Richesson, AK Das, and DL McGuinness. Designing for ai explainability in clinical context. In *Trustworthy AI for Healthcare Workshop at AAAI*, volume 2020, 2021.

[139] Riccardo Guidotti, Anna Monreale, and Leonardo Cariaggi. Investigating neighborhood generation methods for explanations of obscure image classifiers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 55–68. Springer, 2019.

[140] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23, 2019.

[141] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 189–205. Springer, 2019.

[142] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Explaining image classifiers generating exemplars and counter-exemplars from latent representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13665–13668, 2020.

[143] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018.

[144] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[145] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[146] Riccardo Guidotti and Salvatore Ruggieri. On the stability of interpretable models. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[147] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[148] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

[149] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[150] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

[151] Yukinori Harada, Shinichi Katsukura, Ren Kawamura, and Taro Shimizu. Effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: An exploratory analysis of data from a randomized controlled study. *International Journal of Environmental Research and Public Health*, 18(11):5562, 2021.

[152] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[153] Zach Harned, Matthew P Lungren, and Pranav Rajpurkar. Machine vision, medical ai, and malpractice. *Zach Harned, Matthew P. Lungren & Pranav Rajpurkar, Comment, Machine Vision, Medical AI, and Malpractice, Harv. JL & Tech. Dig.(2019)*, 2019.

[154] Kirsten Harrington, Arthur Allen, and Linda Ruchala. Restraining medicare abuse: the case of upcoding. *Research in Healthcare Financial Management*, 11(1):1, 2007.

[155] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[156] Nigel Harvey and Ilan Fischer. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2):117–133, 1997.

[157] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[158] Amy Hawkins. How elderly, sickly farmers are quenching china's thirst for data. *Wired UK*, 2019.

[159] Healthcare Information and Management Systems Society. AI use in European healthcare - HIMSS.

[160] Chadi Helwe, Shady Elbassuoni, Mirabelle Geha, Eveline Hitti, and Carla Makhlouf Obermeyer. Ccs coding of discharge diagnoses via deep neural networks. In *Proceedings of the 2017 International Conference on Digital Health*, pages 175–179, 2017.

[161] Kevin C Heslin, Pamela L Owens, Zeynal Karaca, Marguerite L Barrett, Brian J Moore, and Anne Elixhauser. Trends in opioid-related inpatient stays shifted after the us transitioned to icd-10-cm diagnosis coding in 2015. *Medical care*, 55(11):918–923, 2017.

[162] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. *How humans judge machines*. MIT Press, 2021.

[163] High-Level Expert Group on AI. The Ethics Guidelines for Trustworthy Artificial Intelligence (AI), 04 2019.

[164] Steven D Hillson, Donald P Connelly, and Yuanli Liu. The effects of computer-assisted electrocardiographic interpretation on physicians' diagnostic decisions. *Medical Decision Making*, 15(2):107–112, 1995.

[165] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[166] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[167] Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

[168] Ma Huateng. Application of Artificial Intelligence and Big Data in China's Healthcare Services. In *Global Innovation Index 2019 Creating Healthy Lives - The Future of Medical Innovation*, pages 103–109. 2019.

[169] IBM. Ibm statement on eu ethics guidelines for trustworthy ai.

[170] IBM. IBM's Principles for Trust and Transparency.

[171] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021.

[172] Ergonomics of human-system interaction Human-centred design for interactive systems. Standard, International Organization for Standardization, 2019.

[173] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[174] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):1–9, 2021.

[175] Abhyuday N Jagannatha and Hong Yu. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473. NIH Public Access, 2016.

[176] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[177] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[178] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 360(16):1628–1638, 2009.

[179] Yunzhe Jia, James Bailey, Kotagiri Ramamohanarao, Christopher Leckie, and Michael E Houle. Improving the quality of explanations with local embedding perturbations. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 875–884, 2019.

[180] Zheng Jia, Xudong Lu, Huilong Duan, and Haomin Li. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Medical Informatics and Decision Making*, 19(1):91, 2019.

[181] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

[182] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[183] Bo Jin, Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin, and Xiaopeng Wei. Predicting the risk of heart failure with ehr sequential data modeling. *Ieee Access*, 6:9256–9261, 2018.

[184] Meng Jing and Sarah Dai. China recruits baidu, alibaba and tencent to ai 'national team'. *South China Morning Post*, 2017.

[185] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[186] Ulf Johansson, Lars Niklasson, and Rikard König. Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion*, volume 1, pages 295–300, 2004.

[187] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Mark Roger. Mimic-iv (version 0.4). *PhysioNet*, 2020.

[188] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[189] Hessie Jones. Geoff hinton dismissed the need for explainable AI: 8 experts explain why he's wrong. *Forbes, Dec*, 20, 2018.

[190] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

[191] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[192] Amit Kaushal, Russ Altman, and Curt Langlotz. Geographic distribution of us cohorts used to train deep learning algorithms. *Jama*, 324(12):1212–1213, 2020.

[193] Amit Kaushal, Russ Altman, and Curt Langlotz. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA*, 324(12):1212–1213, 09 2020.

[194] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[195] Kenneth Keppel, Elsie Pamuk, John Lynch, Olivia Carter-Pokras, Insun Kim, Vickie Mays, Jeffrey Pearcy, Victor Schoenbach, and Joel S Weissman. Methodological issues in measuring health disparities. *Vital and health statistics. Series 2, Data evaluation and methods research*, (141):1, 2005.

[196] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics*, 6(2):e24, 2018.

[197] Rafiullah Khan, Arshad Ahmad, Alhuseen Omar Alsayed, Muhammad Binsawad, Muhammad Arshad Islam, and Mohib Ullah. Qupid attack: machine learning-based privacy quantification mechanism for pir protocols in health-related web search. *Scientific Programming*, 2020, 2020.

[198] Seok Kim, Kee-Hyuck Lee, Hee Hwang, and Sooyoung Yoo. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (emr) using the unified theory of acceptance and use of technology (utaut) in a tertiary hospital. *BMC medical informatics and decision making*, 16(1):1–12, 2015.

[199] Young Whan Kim and Jin H Kim. A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation*, 1990.

[200] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[201] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

[202] Ajay Kohli and Saurabh Jha. Why cad failed in mammography. *Journal of the American College of Radiology*, 15(3):535–537, 2018.

[203] C Krittanawong. The rise of artificial intelligence and the uncertain future for physicians. *European journal of internal medicine*, 48:e13–e14, 2018.

[204] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. AIES '20, page 79–85, New York, NY, USA, 2020. Association for Computing Machinery.

[205] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine*, 94:42–53, 2019.

[206] Latrice G Landry, Nadya Ali, David R Williams, Heidi L Rehm, and Vence L Bonham. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, 37(5):780–785, 2018.

[207] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. ICML, 2018.

[208] Frank Lawler, Jim R Cacy, Nancy Viviani, Robert M Hamm, and Stephen W Cobb. Implementation and termination of a computerized medical information system. *Journal of Family Practice*, 42(3):233–236, 1996.

[209] Heidi Ledford. Google health-data scandal spooks researchers. *Nature News*.

[210] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[211] Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. Information retrieval based on conceptual distance in is-a hierarchies. *Journal of documentation*, 1993.

[212] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[213] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[214] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[215] Fengjun Li, Xukai Zou, Peng Liu, and Jake Y Chen. New threats to health data privacy. In *BMC bioinformatics*, volume 12, pages 1–7. BioMed Central, 2011.

[216] He Li, Lu Yu, and Wu He. The impact of gdpr on global technology development, 2019.

[217] Liu Li, Xiaofei Wang, Mai Xu, Hanruo Liu, and Ximeng Chen. Deepgf: Glaucoma forecast using the sequential fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–635. Springer, 2020.

[218] Runzhi Li, Wei Liu, Yusong Lin, Hongling Zhao, and Chaoyang Zhang. An ensemble multilabel classification for disease risk prediction. *Journal of healthcare engineering*, 2017, 2017.

[219] Xin Li and Dongxiao Zhu. Robust detection of adversarial attacks on medical images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1154–1158. IEEE, 2020.

[220] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[221] Thomas Lindow, Josefine Kron, Hans Thulesius, Erik Ljungström, and Olle Pahlm. Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a swedish primary health care setting. *Scandinavian journal of primary health care*, 37(4):426–433, 2019.

[222] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, et al. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596, 2018.

[223] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[224] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[225] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[226] Carissa A Low. Harnessing consumer smartphone and wearable sensors for clinical cancer research. *NPJ Digital Medicine*, 3(1):1–7, 2020.

[227] Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4):303–313, 2012.

[228] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[229] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[230] Sarah M Lyon, Nicole M Benson, Colin R Cooke, Theodore J Iwashyna, Sarah J Ratcliffe, and Jeremy M Kahn. The effect of insurance status on mortality and procedural use in critically ill patients. *American journal of respiratory and critical care medicine*, 184(7):809–815, 2011.

[231] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911. ACM, 2017.

[232] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

[233] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine*, 1(1):1–8, 2018.

[234] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer, 2000.

[235] Gianclaudio Malgieri and Giovanni Comandé. Sensitive-by-distance: quasi-health data in the algorithmic era. *Information & Communications Technology Law*, 26(3):229–249, 2017.

[236] Gianclaudio Malgieri and Giovanni Comandé. Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *Int. Data Privacy Law*, 7(4):243–265, 2017.

[237] Vidushi Marda and Shivangi Narayan. On the importance of ethnographic methods in ai research. *Nature Machine Intelligence*, 3(3):187–189, 2021.

[238] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling. In *2021 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7. IEEE, 2021.

[239] Martijn Millecamp, Sidra Naveed, Katrien Verbert, and Jürgen Ziegler. To explain or not to explain: The effects of personal characteristics when explaining feature-based recommendations in different domains. In *Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, volume 2450, pages 10–18. CEUR; http://ceur-ws. org/Vol-2450/paper2. pdf, 2019.

[240] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[241] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[242] Lorenzo Moja, Hernan Polo Friz, Matteo Capobussi, Koren Kwag, Rita Banzi, Francesca Ruggiero, Marien González-Lorenzo, Elisa G Liberati, Massimo Mangia, Peter Nyberg, et al. Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: A randomized clinical trial. *JAMA network open*, 2(12):e1917094–e1917094, 2019.

[243] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020.

[244] Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. Google Health and the NHS: overcoming the trust deficit. *The Lancet Digital Health*, 1(8):e389, 2019.

[245] Lori Mosca, Elizabeth Barrett-Connor, and Nanette Kass Wenger. Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. *Circulation*, 124(19):2145–2154, 2011.

[246] Annette Moxey, Jane Robertson, David Newby, Isla Hains, Margaret Williamson, and Sallie-Anne Pearson. Computerized clinical decision support for prescribing: provision does not guarantee uptake. *Journal of the American Medical Informatics Association*, 17(1):25–33, 2010.

[247] Paul Mozur, Raymond Zhong, and Aaron Krolik. In coronavirus fight, china gives citizens a color code, with red flags. *The New York Times*, 2020.

[248] Henrik Mucha, Sebastian Robert, Ruediger Breitschwerdt, and Michael Fellmann. Interfaces for explanations in human-ai interaction: Proposing a design evaluation approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[249] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[250] Evan D Muse and Eric J Topol. More than meets the eye: Using ai to identify reduced heart function by electrocardiograms. *Med*, 2(7):791–793, 2021.

[251] Steven Lee Myers, Jin Wu, and Claire Fu. China's looming crisis: A shrinking population. *New York Times*, 2019.

[252] Satheesh Nair, Joseph Eustace, and Paul J Thuluvath. Effect of race on outcome of orthotopic liver transplantation: a cohort study. *The Lancet*, 359(9303):287–293, 2002.

[253] Ju Gang Nam, Sunggyun Park, Eui Jin Hwang, Jong Hyuk Lee, Kwang-Nam Jin, Kun Young Lim, Thienkai Huy Vu, Jae Ho Sohn, Sangheum Hwang, Jin Mo Goo, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*, 290(1):218–228, 2019.

[254] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[255] Emanuele Neri, Francesca Coppola, Vittorio Miele, Corrado Bibbolino, and Roberto Grassi. Artificial intelligence: Who is responsible for the diagnosis?, 2020.

[256] Bret Nestor, Matthew McDermott, Geeticka Chauhan, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. *arXiv preprint arXiv:1811.12583*, 2018.

[257] Crispin Niebel. The impact of the general data protection regulation on innovation and the global political economy. *Computer Law & Security Review*, 40:105523, 2021.

[258] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.

[259] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[260] OECD. Recommendation of the council on artificial intelligence, oecd/legal/0449, 2019.

[261] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[262] High-Level Expert Group on AI. Trustworthy ai assessment list.

[263] World Health Organization et al. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. 2016.

[264] Javier De Velasco Oriol, Edgar E Vallejo, Karol Estrada, José Gerardo Taméz Peña, Alzheimer's Disease Neuroimaging Initiative, et al. Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data. *BMC bioinformatics*, 20(1):1–17, 2019.

[265] Karlijn Overeem, Marjan J Faber, Onyebuchi A Arah, Glyn Elwyn, Kiki MJMH Lombarts, Hub C Wollersheim, and Richard PTM Grol. Doctor performance assessment in daily practise: does it help doctors or not? a systematic review. *Medical education*, 41(11):1039–1049, 2007.

[266] Vijay Pande. Artificial intelligence's' black box'is nothing to fear. *The New York Times*, 2018.

[267] Cecilia Panigutti and Emanuele Bosi. Intelligenza artificiale in ambito diabetologico: prospettive, dalla ricerca di base alle applicazioni cliniche. *Il Diabete Online, Organo ufficiale della Società Italiana di Diabetologia, Medicina traslazionale*, 33(1), 2021.

[268] Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pages 97–110. Springer, 2019.

[269] Cecilia Panigutti, Anna Monreale, Giovanni Comandè, and Dino Pedreschi. Ethical, societal and legal issues in deep learning for healthcare. In *DEEP LEARNING IN BIOLOGY AND MEDICINE*, pages 265–313. World Scientific, 2022.

[270] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657, 2021.

[271] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 629–639, 2020.

[272] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[273] Seong Ho Park and Kyunghwa Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, 2018.

[274] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784, 2019.

[275] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[276] Ignacio Peis, Pablo M Olmos, Constanza Vera-Varela, María Luisa Barrigón, Philippe Courtet, Enrique Baca-García, and Antonio Artés-Rodríguez. Deep sequential models for suicidal ideation from multiple source data. *IEEE journal of biomedical and health informatics*, 23(6):2286–2293, 2019.

[277] Filippo Pesapane, Caterina Volonté, Marina Codari, and Francesco Sardanelli. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in europe and the united states. *Insights into imaging*, 9(5):745–753, 2018.

[278] John P Pestian et al. A shared task involving multi-label classification of clinical free text. In *BioNLP*, pages 97–104. Association for Computational Linguistics, 2007.

[279] Sean M Phelan, Diane J Burgess, Mark W Yeazel, Wendy L Hellerstedt, Joan M Griffin, and Michelle van Ryn. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obesity Reviews*, 16(4):319–326, 2015.

[280] Matthias Pierce, Sally McManus, Curtis Jessop, Ann John, Matthew Hotopf, Tamsin Ford, Stephani Hatch, Simon Wessely, and Kathryn M Abel. Says who? the significance of sampling in mental health surveys during covid-19. *The Lancet Psychiatry*, 7(7):567–568, 2020.

[281] Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[282] Charles Piper. Popular health care provider fraud schemes. *Do No Harm" Isn't Their Motto*, 10.

[283] Mihail Popescu and Mohammad Khalilia. Improving disease prediction using icd-9 ontological features. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 1805–1809. IEEE, 2011.

[284] Mihail Popescu and Dong Xu. *Data mining in biomedicine using ontologies*. Artech House, 2009.

[285] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

[286] Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, and Tadashi Yanagihara. Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. *Transactions on Data Privacy*, 11(2):139–167, 2018.

[287] Michael G Pratt. From the editors: For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research, 2009.

[288] Jennifer Preece, Yvonne Rogers, and Helen Sharp. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, 5 edition, 2019.

[289] Rebecca M Puhl, Joerg Luedicke, and Carlos M Grilo. Obesity bias in training: attitudes, beliefs, and observations among advanced trainees in professional health disciplines. *Obesity*, 22(4):1008–1015, 2014.

[290] Giorgio Quer, Jennifer M Radin, Matteo Gadaleta, Katie Baca-Motes, Lauren Ariniello, Edward Ramos, Vik Kheterpal, Eric J Topol, and Steven R Steinhubl. Wearable sensor data and self-reported symptoms for covid-19 detection. *Nature Medicine*, 27(1):73–77, 2021.

[291] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[292] Jennifer M Radin, Nathan E Wineinger, Eric J Topol, and Steven R Steinhubl. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the usa: a population-based study. *The Lancet Digital Health*, 2(2):e85–e93, 2020.

[293] Parvez Rafi, Arash Pakbin, and Shiva Kumar Pentyala. Interpretable deep learning framework for predicting all-cause 30-day icu readmissions. Technical report, Tech. Rep, 2018.

[294] Sharan Raja and Rudraksh Tuwani. Adversarial attacks against deep learning systems for icd-9 code assignment. *arXiv preprint arXiv:2009.13720*, 2020.

[295] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020.

[296] Zaccharie Ramzi, Philippe Ciuciu, and Jean-Luc Starck. Benchmarking deep nets mri reconstruction models on the fastmri publicly available dataset. In *International Symposium on Biomedical Imaging*, 2020.

[297] Niloofar Rastin, Mansoor Zolghadri Jahromi, and Mohammad Taheri. A generalized weighted distance k-nearest neighbor for multi-label problems. *Pattern Recognition*, page 107526, 2020.

[298] Daniele Ravi, Charence Wong, Benny Lo, and Guang-Zhong Yang. A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics*, 21(1):56–64, 2016.

[299] Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Multi-label classification using ensembles of pruned sets. In *2008 eighth IEEE international conference on data mining*, pages 995–1000. IEEE, 2008.

[300] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[301] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.

[302] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[303] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[304] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. The chinese approach to artificial intelligence: an analysis of policy and regulation. *Available at SSRN 3469784*, 2019.

[305] Kathryn Rough, Andrew M Dai, Kun Zhang, Yuan Xue, Laura M Vardoulakis, Claire Cui, Atul J Butte, Michael D Howell, and Alvin Rajkomar. Predicting inpatient medication orders from electronic health record data. *Clinical Pharmacology & Therapeutics*, 108(1):145–154, 2020.

[306] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[307] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *TKDD*, 4(2):9:1–9:40, 2010.

[308] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–40, 2010.

[309] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[310] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *CoRR*, abs/1811.05577, 2018.

[311] Murali Sambasivan, Pouyan Esmaeilzadeh, Naresh Kumar, and Hossein Nezakati. Intention to adopt clinical decision support systems in a developing country: effect of physician's perceived professional autonomy, involvement and belief: a cross-sectional study. *BMC medical informatics and decision making*, 12(1):1–8, 2012.

[312] Joshua D Schiffman, Paul G Fisher, and Peter Gibbs. Early detection of cancer: past, present, and future. *American Society of Clinical Oncology Educational Book*, 35(1):57–65, 2015.

[313] Philipp Schmidt and Felix Biessmann. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 431–449. Springer, 2020.

[314] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel van den Bosch. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, page 102684, 2021.

[315] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[316] Jessica M Schwartz, Amanda J Moy, Sarah C Rossetti, Noémie Elhadad, and Kenrick D Cato. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. *Journal of the American Medical Informatics Association*, 28(3):653–663, 2021.

[317] Ian A Scott, Ahmad Abdel-Hafez, Michael Barras, and Stephen Canaris. What is needed to mainstream artificial intelligence in health care? *Australian Health Review*, 2021.

[318] Mark Scott. Europe is fighting tech battle with one hand tied behind its back. *Politico*.

[319] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[320] Lesia Semenova and Cynthia Rudin. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

[321] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. The human body is a black box supporting clinical decision-making with deep learning. In *Conference on Fairness, Accountability, and Transparency*, pages 99–109, 2020.

[322] Mark P Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery.

[323] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[324] Mattia Setzu, Riccardo Guidotti, Anna Monreale, and Franco Turini. Global explanations with local scoring. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 159–171. Springer, 2019.

[325] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx-from local to global explanations of black box ai models. *Artificial Intelligence*, page 103457, 2021.

[326] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv:2003.00827*, 2020.

[327] Sharath M Shankaranarayana and Davor Runje. Alime: Autoencoder based approach for local interpretability. In *International conference on intelligent data engineering and automated learning*, pages 454–463. Springer, 2019.

[328] James Shaw, Frank Rudzicz, Trevor Jamieson, and Avi Goldfarb. Artificial intelligence and the implementation challenge. *Journal of medical Internet research*, 21(7):e13659, 2019.

[329] Zhenkun Shi, Weitong Chen, Shining Liang, Wanli Zuo, Lin Yue, and Sen Wang. Deep interpretable mortality model for intensive care unit risk prediction. In *International Conference on Advanced Data Mining and Applications*, pages 617–631. Springer, 2019.

[330] Benjamin Shickel and Parisa Rashidi. Sequential interpretability: Methods, applications, and future direction for understanding deep learning models in the context of sequential data. *arXiv preprint arXiv:2004.12524*, 2020.

[331] Lucy Shinners, Christina Aggar, Sandra Grace, and Stuart Smith. Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: an integrative review. *Health informatics journal*, 26(2):1225–1236, 2020.

[332] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[333] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

[334] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[335] Janet A Sniezek and Timothy Buckley. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes*, 62(2):159–174, 1995.

[336] Janet A Sniezek and Lyn M Van Swol. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes*, 84(2):288–307, 2001.

[337] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[338] Molly Follette Story. Maximizing usability: the principles of universal design. *Assistive technology*, 10(1):4–12, 1998.

[339] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter PC Boon, and Ellen HM Moors. Implementation of artificial intelligence (ai) applications in radiology: hindering and facilitating factors. *European radiology*, 30:5525–5532, 2020.

[340] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[341] Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.

[342] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[343] Hannah R Sullivan and Scott J Schweikart. Are current tort liability doctrines adequate for addressing injury caused by ai? *AMA journal of ethics*, 21(2):160–166, 2019.

[344] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[345] Harini Suresh and John V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle, 2021.

[346] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2007.

[347] Xiaodong Tan, Xiangxiang Liu, and Haiyan Shao. Healthy China 2030: a vision for health care. *Value in health regional issues*, 12:112–114, 2017.

[348] Tencent. Bridging gaps in healthcare industry with technology.

[349] Lancet The. China's health-care reform: an independent evaluation. *Lancet (London, England)*, 394(10204):1113, 2019.

[350] Joseph J Titano, Marcus Badgeley, Javin Schefflein, Margaret Pain, Andres Su, Michael Cai, Nathaniel Swinburne, John Zech, Jun Kim, Joshua Bederson, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature medicine*, 24(9):1337–1341, 2018.

[351] Kevin Tobia, Aileen Nielsen, and Alexander Stremitzer. When does physician use of ai increase liability? *Journal of Nuclear Medicine*, 62(1):17–21, 2021.

[352] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference*, pages 359–380. PMLR, 2019.

[353] Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[354] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[355] Madhukar H Trivedi, JK Kern, A Marcee, B Grannemann, B Kleiber, T Bettinger, KZ Altshuler, and A McClelland. Development and implementation of computerized clinical guidelines: barriers and solutions. *Methods of information in medicine*, 41(05):435–442, 2002.

[356] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[357] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE transactions on knowledge and data engineering*, 23(7):1079–1089, 2010.

[358] Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094911. International Society for Optics and Photonics, 2019.

[359] Helena Varonen, Tiina Kortteisto, Minna Kaila, and EBMeDS Study Group. What may help or hinder the implementation of computerized decision support systems (cdsss): a focus group study with physicians. *Family practice*, 25(3):162–167, 2008.

[360] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[361] Effy Vayena, Anna Mastroianni, and Jeffrey Kahn. Ethical issues in health research with novel online sources. *American Journal of Public Health*, 102(12):2225–2230, 2012.

[362] Siddharth Venkataramakrishnan. Eu backs ai regulation while china and us favour technology. *Financial Times*, 2019.

[363] Viswanath Venkatesh. Adoption and use of ai tools: a research agenda grounded in utaut. *Annals of Operations Research*, pages 1–12, 2021.

[364] Viswanath Venkatesh and Hillol Bala. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2):273–315, 2008.

[365] Viswanath Venkatesh and Fred D Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204, 2000.

[366] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

[367] Viswanath Venkatesh, Tracy A. Sykes, and Xiaojun Zhang. 'just what the doctor ordered': A revised utaut for emr system adoption and use by doctors. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10, 2011.

[368] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73(2):185, 2008.

[369] Himanshu Verma, Roger Schaer, Julien Reichenbach, Jreige Mario, John O Prior, Florian Evéquoz, and Adrien Raphaël Depeursinge. On improving physicians' trust in ai: Qualitative inquiry with imaging experts in the oncological domain. 2021.

[370] Alexandr Vesselkov, Heikki Hämmäinen, and Juuso Töyli. Design and governance of mhealth data sharing. *Communications of the Association for Information Systems*, 45(1):18, 2019.

[371] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10, 2020.

[372] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: obtaining reliable explanations for machine learning models. *arXiv preprint arXiv:2001.11757*, 2020.

[373] Amy Walker. Nhs gives amazon free use of health data under alexa advice deal. *The Guardian*, 2018.

[374] Shirly Wang, Matthew McDermott, Geeticka Chauhan, Michael C Hughes, Tristan Naumann, and Marzyeh Ghassemi. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. *arXiv preprint arXiv:1907.08322*, 2019.

[375] Shui-Hua Wang, Vishnu Govindaraj, Juan Manuel Gorriz, Xin Zhang, and Yu-Dong Zhang. Explainable diagnosis of secondary pulmonary tuberculosis by graph rank-based average pooling neural network. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–14, 2021.

[376] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1):2357–2393, 2017.

[377] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328. Association for Computing Machinery, 2021.

[378] Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania. Full translation: China's 'new generation artificial intelligence development plan'(2017). *DigiChina, August*, 1, 2017.

[379] Wei Weng, Yu-Wen Li, Jing-Hua Liu, Shun-Xiang Wu, and Chin-Ling Chen. Multi-label classification review and opportunities. 2021.

[380] Kathleen McGlone West, Erika Blacksher, and Wylie Burke. Genomics, health disparities, and missed opportunities for the nation's research agenda. *Jama*, 317(18):1831–1832, 2017.

[381] KA Wetterstrand. NA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP).

[382] White House's Office of Science and Technology Policy. Guidance for regulation of artificial intelligence applications, 2019.

[383] World Health Organization WHO et al. Icd purpose and uses. *Classification. Available online at: http://www.who. int/classifications/icd/en/(Accessed May 20, 2020)*, 2018.

[384] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[385] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care (vol 25, pg 1337, 2019). *Nature Medicine*, 25(10):1627–1627, 2019.

[386] Hannah B Wild. There's no algorithm for empathy. *Health Affairs*, 39(2):339–342, 2020.

[387] Scott Wisdom, Thomas Powers, James Pitton, and Les Atlas. Interpretable recurrent neural networks using sequential sparse recovery. *arXiv preprint arXiv:1611.07252*, 2016.

[388] Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584, 2021.

[389] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[390] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

[391] Peizhen Xie, Ke Zuo, Yu Zhang, Fangfang Li, Mingzhu Yin, and Kai Lu. Interpretable classification from skin cancer histology slides using deep learning: A retrospective multicenter study. *arXiv preprint arXiv:1904.06156*, 2019.

[392] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[393] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573. ACM, 2018.

[394] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs): a survey. *ACM Computing Surveys (CSUR)*, 50(6):85, 2018.

[395] Jiamin Yin, Kee Yuan Ngiam, and Hock Hai Teo. Role of artificial intelligence applications in real-life clinical practice: Systematic review. *Journal of medical Internet research*, 23(4):e25759, 2021.

[396] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 223–227, 2016.

[397] Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

[398] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, number 3, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[399] Henry Zhan, Kevin Schartz, Matthew E Zygmont, Jamlik-Omari Johnson, and Elizabeth A Krupinski. The impact of fatigue on complex ct case interpretation by radiology residents. *Academic radiology*, 28(3):424–432, 2021.

[400] Boyu Zhang, Anis Zaman, Vincent Silenzio, Henry Kautz, and Ehsan Hoque. The relationships of deteriorating depression and anxiety with longitudinal behavioral changes in google and youtube use during covid-19: Observational study. *JMIR Mental Health*, 7(11):e24012, 2020.

[401] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

[402] Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008, 2010.

[403] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[404] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[405] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[406] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8), 2014.

[407] Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 249–256, 2020.

[408] Ping Zhang and Yuan Liang. China's national health guiding principles: a perspective worthy of healthcare reform. *Primary health care research & development*, 19(1):99–104, 2018.

[409] Xiaoqing Zhang, Hongling Zhao, Shuo Zhang, and Runzhi Li. A novel deep neural network model for multi-label chronic disease prediction. *Frontiers in genetics*, 10:351, 2019.

[410] Xuezhou Zhang, Sarah Tan, Paul Koch, Yin Lou, Urszula Chajewska, and Rich Caruana. Interpretability is harder in the multiclass setting: axiomatic interpretability for multiclass additive models. *age*, 25(50):75–100, 2019.

[411] Yudong Zhang, Xin Zhang, and Weiguo Zhu. Anc: Attention network for covid-19 explainable diagnosis based on convolutional block attention module. *Computer Modeling in Engineering & Sciences*, pages 1037–1058, 2021.

[412] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[413] Wenting Zhao, Shufeng Kong, Junwen Bai, Daniel Fink, and Carla Gomes. Hot-vae: Learning high-order label correlation for multi-label classification via attention-based variational autoencoders. *arXiv preprint arXiv:2103.06375*, 2021.