Co-design of human-centered, explainable AI for clinical decision support

CECILIA PANIGUTTI, Università di Pisa, Italy and European Commission, Joint Research Centre (JRC), Italy
ANDREA BERETTA, CNR, Italy
DANIELE FADDA, CNR, Italy
FOSCA GIANNOTTI, CNR, Italy and Scuola Normale Superiore, Italy
DINO PEDRESCHI, Università di Pisa, Italy
ALAN PEROTTI, CENTAI Institute, Italy
SALVATORE RINZIVILLO, CNR, Italy

eXplainable AI (XAI) involves two intertwined but separate challenges: the development of techniques to extract explanations from black-box AI models, and the way such explanations are presented to users, i.e., the explanation user interface. Despite its importance, the second aspect has received limited attention so far in the literature. Effective AI explanation interfaces are fundamental for allowing human decision-makers to take advantage and oversee high-risk AI systems effectively. Following an iterative design approach, we present the first cycle of prototyping-testing-redesigning of an explainable AI technique, and its explanation user interface for clinical Decision Support Systems (DSS). We first present an XAI technique that meets the technical requirements of the healthcare domain: sequential, ontology-linked patient data, and multi-label classification tasks. We demonstrate its applicability to explain a clinical DSS, and we design a first prototype of an explanation user interface. Next, we test such a prototype with healthcare providers and collect their feedback, with a two-fold outcome: first, we obtain evidence that explanations increase users' trust in the XAI system, and second, we obtain useful insights on the perceived deficiencies of their interaction with the system, so that we can re-design a better, more human-centered explanation interface.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**; **User studies**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: explainable artificial intelligence, clinical decision support systems, human-computer interaction, user study

1 INTRODUCTION

Many aspects of our lives, including our health, have become increasingly digitalized over the last decade. Healthcare data can be collected from various sources, including Electronic Health Records (EHRs) maintained by healthcare providers [68], health and wellness apps [61, 105], and wearable sensors [27, 104]. The wealth of information contained in such data has the potential to provide valuable insights into an individual's health,

*The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

Authors' addresses: Cecilia Panigutti, cecilia.panigutti@ec.europa.eu, Università di Pisa, Pisa, Italy and European Commission, Joint Research Centre (JRC) , Ispra, Italy; Andrea Beretta, andrea.beretta@isti.cnr.it, CNR, Pisa, Italy; Daniele Fadda, daniele.fadda@isti.cnr.it, CNR, Pisa, Italy; Fosca Giannotti, fosca.giannotti@sns.it, CNR, Pisa, Italy and Scuola Normale Superiore, Pisa, Italy; Dino Pedreschi, dino.pedreschi@unipi.it, Università di Pisa, Pisa, Italy; Alan Perotti, alan.perotti@centai.eu, CENTAI Institute, Turin, Italy; Salvatore Rinzivillo, rinzivillo@isti.cnr.it, CNR, Pisa, Italy.

2160-6455/2023/3-ART
https://doi.org/10.1145/3587271

potentially in real-time, and could be used to improve healthcare delivery [5, 9, 67, 75, 115]. However, this *data deluge* can be overwhelming for humans to analyze and interpret, leading to a need for new and improved data processing methods. Artificial Intelligence (AI) can be used to identify patterns and trends in such large datasets and has the potential to help healthcare professionals make more informed and efficient decisions about patients care [30, 128]. One approach for integrating AI into clinical practice is using it in clinical Decision Support Systems (DSS), i.e., computerized systems that give evidence-based suggestions, alerts, and reminders to healthcare practitioners to help with patient diagnosis, treatment, and management [126]. However, the rate of adoption of AI in health clinics and hospitals is low [4, 114, 140]. A recent report estimates that 84% of healthcare providers in Europe currently do not use any AI system [62]. The reasons behind the low adoption of clinical DSS that do not embed AI have been well studied, with shortcomings including perceived challenges to autonomy, lack of time, and dissatisfaction with user interfaces [17, 72, 74, 78, 94, 110, 129, 130]. In addition to these adoption barriers, AI-based clinical DSS also face trust issues from medical staff and a lack of knowledge about the assumptions, limitations and capabilities of such systems [62, 131]. 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 [49, 108, 134, 139]. The study of techniques whose goal is to explain (i.e., capability to present in human-understandable terms [43]) the decision-making process of an AI system is the main focus of the eXplainable AI (XAI) field of research. This topic has recently witnessed an increased interest that generated vast literature on AI transparency and explainability [19, 57]. 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. 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 [93, 101].

Furthermore, XAI techniques could help achieve the highest levels of AI transparency for high-risk AI applications (such as healthcare) as mandated by the AI Act, a recent EU regulation proposal on AI [2, 122]. Indeed, the AI Act prescribes that high-risk AI systems should be designed with human-machine interfaces that enable users to interpret the system's output and use it appropriately. Academic literature has also debated the existence of a "right to have an explanation" of AI decisions based on the EU GDPR (General Data Protection Regulation) prescription to provide "meaningful information about the logic involved" to the data subject in case of automated decision-making [1, 36, 58, 87]. 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 [66]. Furthermore, explanations might even increase overreliance on AI-based clinical DSS [23, 47, 76], 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 [22]. These findings highlight the importance of involving the end-user of the explanation in the evaluation or, ideally, already in the design phase. A recent survey has argued that explanations of black-box AI models are mainly used by machine learning engineers to debug their model in the development phase [15]. Nevertheless, debugging the model is only one of the needs expressed in another recent study that analyzes the demands of transparency of several stakeholders [20]. The fact that the developers of XAI methods design explanations for themselves creates a gap between state-of-the-art XAI explanations and end-users, undermining the impact of XAI in high-stakes decision-making.

In this paper, we present the collective effort of our interdisciplinary team of data scientists, human-computer interaction experts and designers to develop a human-centered, explainable AI system for clinical decision support. Using an iterative design approach that involves healthcare providers as end-users, we present the first cycle of the prototyping-testing-redesigning of the explainable AI technique and its explanation user interface. We first present the XAI technique's conception that stems from patients data and healthcare application requirements. Then we develop the initial prototype of the explanation user interface, and perform a user study to test its perceived trustworthiness and collect healthcare providers' feedback. We finally exploit the users' feedback to

co-design a more human-centered XAI user interface taking into account design principles such as progressive disclosure of information.

Co-design is a participatory design approach that involves the end users as active participants in the design process [106]. This approach to design is particularly useful when developing human-centered XAI interfaces whose goal is to facilitate smooth and useful human-machine interactions. When considering user involvement in the design process of an AI explanation, it is important to disentangle two aspects of XAI development: the technical development of a technique able to extract an explanation from the black-box AI model (XAI), and the way such explanation is presented to users, i.e. the explanation user interface (XUI) [34]. Since developing an XAI methodology is an extremely technical endeavor, we did not involve users in this first stage of development.

We build upon our previous work *Doctor XAI* [103]. We add new mechanisms to tailor Doctor XAI towards a clinical prediction task and we carry out a thorough technical evaluation of the methodology, demonstrating its applicability to the explanation of a clinical DSS. We then translate the rule-based explanation of Doctor XAI in natural language and present the first prototype of a user explanation interface. This prototype is then evaluated by healthcare providers via an online user study, obtaining two important insights. First, we get evidence that explanations increase users' trust in the XAI system, and second, we obtain useful insights on the perceived deficiencies of their interaction with the system, so that we can re-design a better, more human-centered explanation interface. Design and evaluation of an explainable user interface are iterative, intertwined processes: the evaluation allows to discover new requirements for the explanation, which is then redesigned to fulfill those requirements and re-evaluated. Our study shows that co-designing XAI-based tools for high-stakes decision-making, such as in healthcare, opens up opportunities for much more trustworthy and responsible use of AI in high-risk tasks. In principle, XAI may be the key for a synergistic human-machine interaction and collaboration, where the uniquely human capabilities are enhanced by the AI's. At the same time, more empirical and theoretical research is needed to better understand the risks associated with over-reliance and automation bias that might stem from advanced XAI tools.

The paper is structured as follows. In section 2, we briefly discuss the main related works in the field of XAI and HCI. In section 3, we present the technical aspects of the XAI methodology and how it meets functional and data requirements associated with healthcare settings. In section 4, we show a technical validation of the XAI technique and its applicability to the explanation of a clinical DSS, demonstrating its ability to accurately represent the black-box model decision-making process. In section 5, we introduce the first explanation user interface prototype. In section 6, we discuss the results of our first user study that explores the relationship between trust in the clinical DSS and AI explanations. This user study also provides insights related to explanations relevancy, i.e. the explanation ability to *provide insights for a particular audience into a chosen domain problem* [96]. Finally in section 7 we present a refinement of the design of the AI explanation interface in response to user feedback.

2 RELATED WORK

While several XAI methods have been developed in the past years, only a few considered the specific application domain. Consider, for example, two of the most popular XAI methods: LIME [108] and SHAP [86]. Similar to the XAI method presented in this paper, they provide local explanations that summarize each feature's influence on the model outcome [38]. These two methods are *model-agnostic* and *application-agnostic*, meaning that they are able to extract an explanation from any type of black-box AI model [91] regardless of the application domain. While the *model-agnostic* approach to XAI offers great flexibility to the use of these methods, the *application-agnostic* approach implies that the specific user needs are not considered [8]. Our XAI methodology fit in an emerging line of research focusing on XAI techniques that are not completely agnostic and tailor their explanations to the medical field, either by incorporating medical knowledge in the explanation process [11, 32, 141] or focusing on specific healthcare data characteristics and use cases [89, 100, 102]. Even though this line of research is a first step

in the direction of considering healthcare professionals' needs, these methods rarely design the explanation with the end-user in mind (a notable exception is [117] where clinicians are involved throughout the development process of the AI application). Furthermore, only a few of them tested the efficacy of their explanations on a group of health care professionals.

A few works have tried to close such a gap in the medical field by involving the doctors in the design procedure [77, 113, 139] or by performing exploratory surveys [26, 83, 127]. Despite these recent efforts, most of the research has been focused on laypeople [7, 29, 95]. However, several works have shown that users' domain expertise is relevant to the trust calibration process [52, 99, 135, 145], 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 [16, 21, 45, 88], i.e., while performing a task supported by a clinical DSS.

Ideally, explaining clinical DSS recommendations should help clinicians with *trust calibration* [73], i.e., properly adjusting their level of trust according to the actual reliability of the AI system [111]. 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 [81] or accepting incorrect diagnoses suggested by AI [59]. This phenomenon is known as *automation bias* [79, 118] 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 [26, 116] and might be subject to *algorithm aversion* [40], which is the human tendency to discount algorithmic advice [85]. Distrust in AI applications in medicine also comes from doctors' fear of legal repercussions if something goes wrong due to unclear liability regimes [97, 124].

This paper presents a user study that investigates how AI explanations impact users' trust in algorithmic recommendations in the healthcare context. Our experimental design allows understanding whether AI explanations increase trust in the AI suggestions in the event that the suggestion is correct. Such an experimental setting offers some initial insights into how AI explanation might influence the trust calibration process. We also use participants' feedback on the explanation interface to improve its design and increase the AI system usability and trust by adopting a co-design approach to improve the usability of such interface. The user interface conveys the transparency of the AI system [73] and, with the advent of XAI, the system explains its recommendations through the interface [10]. The quality of such an explanation interface is of pivotal importance and its design needs to be studied using an HCI lens [46]. Indeed, while the interface of an AI model does not influence its capabilities, early works have shown the relationship between users' trust and the interface design. Indeed, the design of the AI interface can influence users' beliefs about its capabilities and trust in its decisions [79]. It was also proved that some interface features such as its ease of use and its usefulness can increase users' trust in automation [63] and that explanation interface design choices such as the *progressive disclosure* of information can help users' decision-making process [22].

3 THE EXPLAINER: DOCTOR XAI

In this section, we introduce the XAI technique employed in our use case: *Doctor XAI* [103]. Doctor XAI is an explainability technique tailored to medical AI applications. It is *model-agnostic*, i.e., it does not use any internal parameter of the model to generate its explanations, and its explanations are *local*, i.e., they explain the rationale behind the classification of a single data point. Its explanations are a good representative of a common type of AI explanation: the *removal-based type* of explanation [37]. Like other popular removal-based approaches, Doctor XAI explanations summarize each feature's influence on the model outcome [86, 108]. However, unlike other removal-based approaches, it also employs medical knowledge (in the form of medical ontologies) in the explanation extraction process, meaning that the features highlighted to be important were selected considering

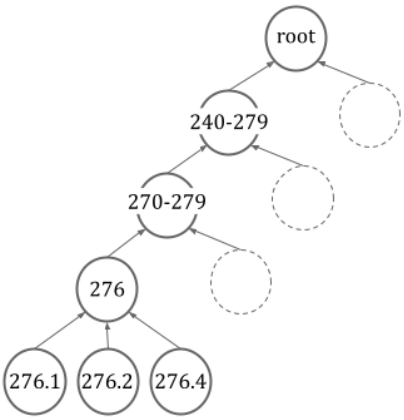
their medical meaning. Furthermore, Doctor XAI is also designed to deal with some typical healthcare data characteristics: sequentiality and multi-label predictions. The following paragraphs illustrate why these data characteristics are relevant to the healthcare context and how we formalized them.

Ontology-linked data. Ontologies (also known as knowledge graphs) are structured, machine-readable representations of the knowledge pertaining to a specific aspect of a domain [56, 125]. An ontology defines a set of vocabulary terms of the domain of interest, also referred to as *concepts*, and encodes all their relevant properties and relationships. 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*, their items can be linked to the concepts represented in an ontology. 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 (DO) [112], the Open Biomedical Ontology (OBO) [119], the Diabetes mellitus Diagnosis Ontology (DDO) [48], the Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) [42] and the Unified Medical Language System (UMLS) [18]. In our use case, we focus on the *International Classification of Diseases* (ICD) ontology. The ICD is the standard for the reporting and coding of diseases and health conditions [137]. In its Ninth Revision, Clinical Modification (ICD-9-CM), the codes have an alphabetic or numeric first digit, and 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 [28]:

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

So the first three digits identify the category of the diagnosis (e.g., *infectious and parasitic diseases*, *endocrine, nutritional and metabolic diseases, and immunity disorders*), while the last two digits identify the etiology or the anatomic site of the diagnosis. The set of hierarchical relationships between these codes constitutes an ontology containing concepts related by the simple relationship type *"is-a"*, e.g. *276.1: Acidosis* "is-a" *276: Diseases of fluid electrolyte and acid-based balance* "is-a" *270-279: Other Metabolic Diseases And Immunity Disorders* "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 1, first plot.

Sequentiality. Sequential data, i.e., any data that contains instances whose representation implies some sort of order, are 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. In our use case, we considered an AI model that processes sequential data tracking patients' hospital encounters over many years, i.e., longitudinal data containing patients' clinical history. In particular, each hospital encounter (or visit) is associated with a list of ICD-9 codes that encode the relevant conditions treated during the hospital stay, making it *sequential* and *ontology-linked* data. More formally, if we define the set of ICD-9 codes as $C = \{c_1, c_2, \ldots, c_{|C|}\}$, each patient's clinical history is represented by a sequence of visits $V_1, \ldots, V_M$ such that each V_i, is a sequence of ICD-9 codes $\subseteq C$. We remark that, while the order of visit is always preserved and factored in our pipeline, the order of clinical codes within each visit is of no importance. A simple example of a patient clinical history representation is shown in figure 1, second plot. The patient visited the hospital three times. During his first stay, he was diagnosed with *hyposmolality* (ICD-9 code 276.1) and *acidosis* (ICD-9 code 276.2), during the second stay he was diagnosed with a *diseases of esophagus* (ICD-9 code



ICD-9	Description
root	Disease
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

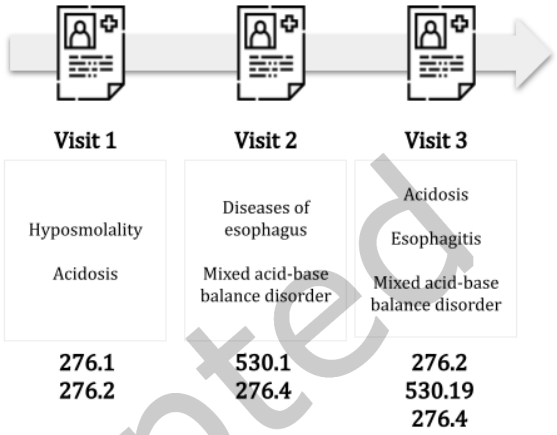


Fig. 1. *(1st plot)* A representation of a branch of the tree-shaped ICD-9 hierarchical ontology: the root is a general condition *Disease* while its children and grandchildren are increasingly more specific conditions. *(2nd plot).* A simple example of the representation a patient clinical history as a sequence of ICD-9 codes.

530.1) and with a *mixed acid-base balance disorder* (ICD-9 code 276.4). Finally in his third and last visit to the hospital he was diagnosed again with *acidosis* and *mixed acid-base balance disorder* and a more specific diseases of esophagus, *esophagitis* (ICD-9 code 530.19). The final representation of the clinical history as a sequence of ICD-9 codes is as follows:

$$[[276.1, 276.2], [530.1, 276.4], [276.2, 276.4, 530.19]]$$

Multi-label predictions. Multi-label prediction tasks, i.e., learning to assign a set of non-mutually exclusive labels to each instance, are often encountered in AI healthcare applications. For example, when there is the need to simultaneously predict the risk of several chronic diseases [50, 51, 82, 144], when trying to classify unknown genes functional expressions [12, 35], when building a clinical algorithm to predict the diagnoses and medications order of patient's future visit [31, 109], when trying to learn multiple indicators of early-stage diseases [33] or when performing clinical text categorization or annotation [13, 41, 142]. In our use case, we considered an AI model that perform a multi-label *next-step prediction*, i.e. it predicts all the diagnoses that will be associated with a patient next visit to the hospital.

Even if Doctor XAI deals *by-design* with sequential, multi-labeled and ontology-based data, none of these features is strictly necessary. Indeed, this explainer can be used with datasets displaying any combination of the three aforementioned features, by exploiting only the corresponding specific modules. We will show in the technical validation section (section 4) the performance of Doctor XAI with and without ontology-linked data

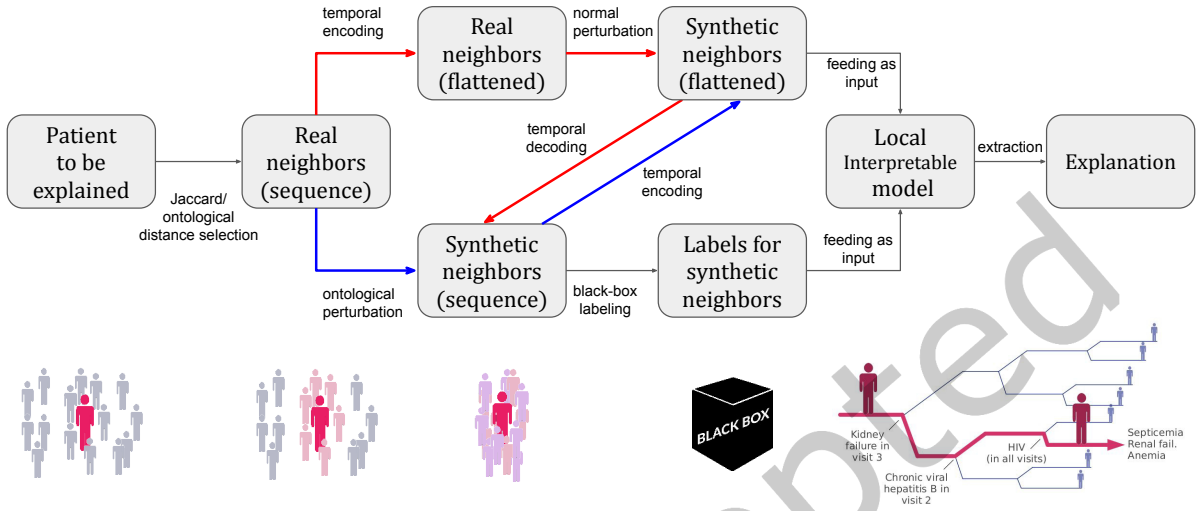


Fig. 2. The explanation pipeline in the case of ontolology-liked data (blue) and not ontology-linked data (red)

and on multilabel and binary classification tasks. Next section illustrates the components of Doctor XAI and how they form the full explanation pipeline.

3.1 Doctor XAI explanation pipeline

Doctor XAI is based on the idea presented in [108] of 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. More formally:

Given an instance x and its black-box outcome $b(x) = y$, an explanation is extracted for this individual decision from an inherently interpretable model c trained to mimic the local behavior of b.

An overview of the methodology is presented in figure 2. The starting point is the data point whose black-box prediction we are interested in explaining. As the first step, Doctor XAI selects 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. These neighbors can be either selected according to a standard distance metric, such as the Jaccard one or exploit ontology-based similarities if the data is ontology-linked. The latter case is described in section 3.2. In both cases, a set of k real neighbors is obtained, each of which is represented as a sequence.

Doctor XAI then generates the synthetic neighborhood perturbing the real neighbors to ensure the *locality* of the resulting 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. In the case of ontology-linked data, Doctor XAI uses the domain knowledge encoded

in the ontology to generate meaningful synthetic instances, as explained in section 3.3. 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 a dense training set for the interpretable classifier c increases its performance in mimicking the black-box [39]. Unlike other explanation techniques, Doctor XAI does not perturb directly the features of the instance whose black-box decision we want to explain. By doing so, it prevents 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, Doctor XAI ensures the *expressiveness* of the synthetic neighborhood, i.e., the black-box classifications are heterogeneous among the synthetic neighbors.

For the perturbation steps in the Doctor XAI pipeline, it is possible to follow two alternative paths, represented by the red and blue arrows in figure 2 (the two paths share the black arrows). The red path is for data that are not ontology-linked and it involves a normal perturbation of the real neighbors, described in section 3.4. The blue path is for ontology-liked data and it involves the ontological perturbation of real neighbors, as described in section 3.3. Both paths involve steps of temporal encoding/decoding (with the relative algorithms described in section 3.5), since the black-box model requires a sequential input, whereas the interpretable one requires a tabular (flat) one.

The red path involves the normal perturbation of the real neighbors: first, they are encoded (flattened) into sparse vectors. Then the normal perturbation is applied in order to obtain a synthetic neighborhood. In order to obtain the labels for the synthetic data points, however, Doctor XAI has to decode them back into sequences so that they can be fed into the black-box model for labeling.

Once Doctor XAI generated both the synthetic neighborhood and the corresponding labels, it can train the interpretable model, and finally, extract symbolic rules. Similarly to [100], Doctor XAI uses a a multi-label decision tree as the inherently interpretable classifier c, however, in the case of binary predictions the decision tree is a binary one. From such decision tree, Doctor XAI extracts rule-based explanations in the form $p \rightarrow y$ where $y = c(x)$. Such 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, Doctor XAI applies the perturbation directly on sequential data, obtains a synthetic neighborhood as a set of sequences, and feeds them to the black-box model for labeling. However, as it was for the red path, the interpretable model requires a tabular input, so Doctor XAI proceeds 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 Doctor XAI follows a general framework for its model-agnostic explanation pipeline, it extends the framework with novel contributions in order to deal with structured data and sequential data respectively. These components can be independently plugged in an explanation pipeline according to the nature of the data point to be explained.

3.2 Ontological distances

In this section, we define a new distance measure that allows Doctor XAI to select the closest real neighbors of the instance whose decision we want to explain. As already mentioned in section 3, each patient's clinical history is represented as a list of visits, which in turn are represented as lists of ICD-9 codes. Every instance is therefore a list of lists of ICD-9 codes. We observe that to performa a multi-hot encoding of 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) [138] because it is one of the most commonly used for ICD-9 ontologies [6, 54, 69].

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 is defined as:

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

$WuP(c_1, c_2) \in [0, 1]$ 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. Indeed visits are defined as lists of occurring ICD-9 codes. We adopted the weighted Levenshtein [80] 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 [14], 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.

3.3 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 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

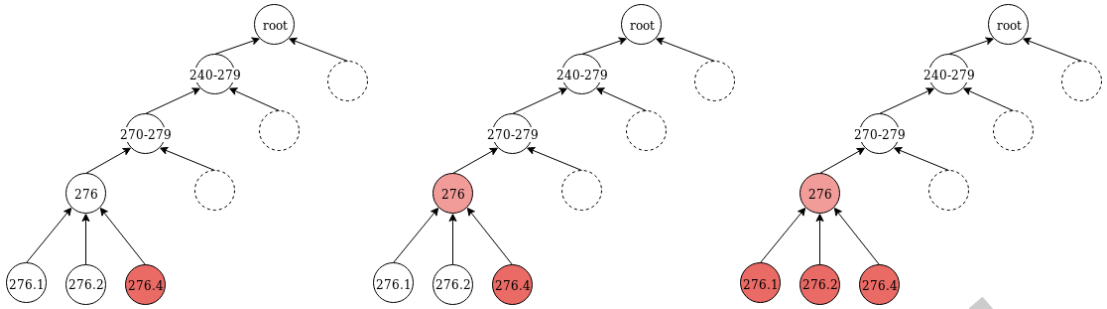


Fig. 3. *(1st plot)* The node corresponding to the randomly selected ICD-9 code (276.4) of the patient is highlighted in red in the ICD-9 ontology graph representation. *(2nd plot).* The ontological superconcept of the selected ICD-9 is selected and highlighted (276). *(3rd plot)* All ICD-9 codes all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

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 1, first plot. 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:

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

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 3. 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.

3.4 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. While such an assumption is not

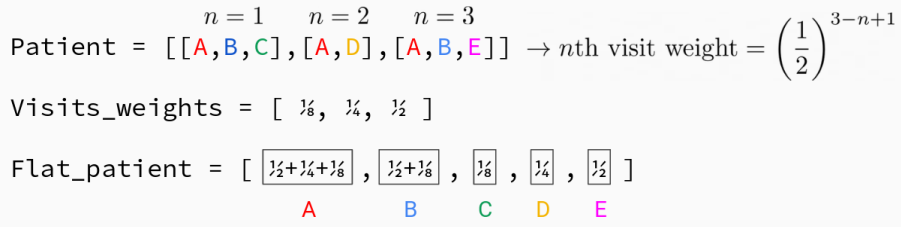


Fig. 4. Example of temporal encoding for a patient

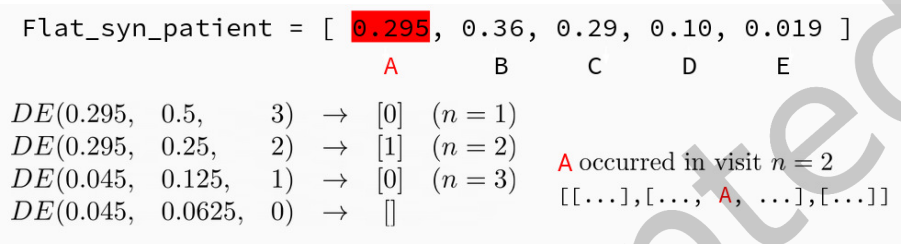


Fig. 5. Example of temporal decoding for a patient

realistic, many popular XAI methodologies (such as LIME [108] on tabular data) use this assumption in their perturbation phase.

3.5 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 4. 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. 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.

4 TECHNICAL VALIDATION OF THE EXPLAINER

We now illustrate the technical validation of the proposed explainer, Doctor XAI. Similarly to many other explainers, Doctor XAI is based on the assumption that an interpretable classifier can mimic the local decision boundary of the black-box model, therefore the metrics used to evaluate its performance must investigate whether this is true. In particular, we chose three metrics to perform such technical evaluation:[1] *fidelity*, *hit*, and *explanation complexity*:

- *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. In the case of multi-label classification, the fidelity is calculated using the F_1 measure with micro-averaging [143], while in the binary case the fidelity is calculated using the accuracy score.
- *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. In the case of multi-label classification, the hit is calculated in the following way: $1 -$ hamming-distance(y_b, y_c), whereas in the binary case it is calculated as $1 - (y_b - y_c)$.
- *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 [43, 84].

We chose to compare the performance of Doctor XAI according to these metrics for different purposes:

- First, we wanted to validate that increasing the local density through the generation of synthetic neighbors increases the fidelity and hit of the local interpretable model.
- Second, we wanted to validate that, in the case of ontology-linked data, exploiting the ontology in the explanation process (in the selection of first neighbors and the synthetic neighborhood generation) increases the fidelity and hit of the local interpretable model.
- Finally, we wanted to optimize the trade off between explanation complexity and fidelity. Indeed a more complex rule allows a more precise characterization of the decision boundary of the black box. At the same

[1]The system's response time while making an explanation, among other metrics, might play an important role in its usability and therefore is relevant to this paper. However, it was not considered in our current investigation as explanations for the user study were not computed in real-time. Future research will consider this and other metrics.

time, a complex rule is less interpretable than a shorter one. We then investigated the ideal number k of real neighbors to consider to optimize both fidelity and explanation complexity.

We performed these experiments both in the case of a multi-label and binary prediction task.

4.1 Experimental setup

This section illustrate the experimental setup used to perform the technical validation of the explainer.

Dataset: MIMIC-IV. The MIMIC (Medical Information Mart for Intensive Care) [55, 71] database is a single-center, publicly accessible database that contains de-identified clinical data from patients admitted to the Beth Israel Deaconess Medical Center's ICU (intensive care unit). MIMIC-IV [70], the most recent version, comprises data on 383220 patients who were admitted to hospitals between 2008 and 2019, totaling 524520 admissions. Each admission's demographics, clinical measures, diagnoses, and procedures codes are stored in the database. The MIMIC database is often used by the Machine Learning research community to train benchmark AI models on freely accessible data to ensure the reproducibility of results [107]. We selected a subset of patients having at least two ICD-9-coded hospital admissions, limiting the number of patients to 43697 and the number of hospital admissions to 164411 (see table 1).

number of patients	43697
number of admissions	164411
avg. nr. of admissions per patient	3.76
max nr. of admissions per patient	146
number of unique ICD-9 codes	8259
avg. nr. of codes per admission	11.22

Table 1. MMIC-IV: Data from patients with at least two hospital admissions

Black box: Doctor AI. *Doctor AI* [31] 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, leaving aside the prediction of medication order and time duration until the next visit. In this simplified setting, the black box does not use any information on the time difference between two visits. 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 is evaluated using *recall@n* with $n = 10, 20, 30$ achieving 0.79 *recall@30* on the private dataset used in the original paper. We pre-processed the MIMIC-IV dataset for Doctor AI following the pre-processing script available in Doctor AI GitHub repository[2]. We then split the dataset in training (29714 patients, 65%) validation (5244 patients, 12%) and test set (8739 patients, 20%) and we trained Doctor AI for 40 epochs using default hyperparameters. Doctor AI can either be trained to forecast a patient's future clinical event in terms ICD codes or CCS (Clinical Classifications Software) categories[3]. CCS categories are used to group ICD-9 codes into a smaller number of clinically relevant groups (e.g. CCS 100 *acute myocardial infarction* groups all the ICD-9 codes related to that condition). We trained Doctor AI to predict which set of CCS categories will be associated with future visit at time $t + 1$ given the patient's ICD-9 clinical history up to visit t. By doing so, we reduced the dimensionality of the label space from 8259 ICD-9 codes to 273 CCS codes. The resulting performace of Doctor AI on the test set are in line with those of the original paper *recall@10* = 0.48, *recall@20* = 0.62 and *recall@30* = 0.70.

[2]https://github.com/mp2893/doctorai

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

Explanation evaluation. We evaluated the explanations for a random subset of 1000 patients of the test test suffering from common heart conditions (CCS categories ranging from 96 to 118[4]). This random sampling was necessary for computational reasons. For each of these patients we evaluated the fidelity and hit of the local model to the black box under several conditions.

- **Exploiting the ontology or not** We compared the explanation pipeline that exploits the knowledge encoded into the ICD-9 ontology to create the synthetic neighborhood (blue path in figure 2) to the pipeline that does not use this semantic information (red path in figure 2). While the ontological explanation pipeline selects the first k real neighbors using the *ontological distance* described in section 3.2 and then generates the synthetic neighborhood by perturbing them using the *ontological perturbations* described in section 3.3, the other pipeline selects the k real neighbors using the *Jaccard similarity* then perturbs them by using *normal perturbations* 3.4.
- **Real or synthetic neighborhood** For each patient we trained two decision trees. One was trained directly on the k real neighbors of that patient from the dataset, while the other one was 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.
- **Binary or multi-label** We carried out experiments in the case of a multi-label and binary explanation task. For the binary explanation task we selected all the patients predicted by the black box as having an *acute mycordial infarction* (CCS 100, 33 patients) and *pulmonary heart disease* (CCS 103, 122 patients).

The code to run our experiments as well as our results are available on GitHub[5].

4.2 Results

Multi-label classification task. We first evaluated the explainer fidelity to the black box for the multi-label classification task. The results are show in figure 6 where we show the fidelity sample distributions at different values of k for the decision trees trained directly on the first k neighbors selected with the Jaccard (blue boxplot) and ontological (greed boxplot) distance, and the decision trees trained on synthetic neighbors generated with the normal (yellow boxplot) and ontological perturbation (red boxplot). 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 explanation pipeline that use the ontology and that which does not. 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 information (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. Finally we can see that the fidelity to the black box decreases in all conditions when the number of first k neighbors is increased. This trend can be observed in the mean fidelity values reported in table 2 (left side), and it also confirmed by the results of the Tukey's HSD test for multiple comparisons shown in the right side of table 2. These tests found that the mean value of fidelity was significantly different between $k = 10$ and $k = 30$ and between $k = 10$ and $k = 20$ for all types of neighborhood, while there was no statistically significant difference between $k = 20$ and $k = 30$ for the real and synthetic neighborhood created without the ontological information. These results, together with the fact that fidelity is higher for lower values of k, indicate that the local multi-label decision tree struggles to approximate the decision

[4]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf
[5]https://github.com/CeciPani/DrXAI

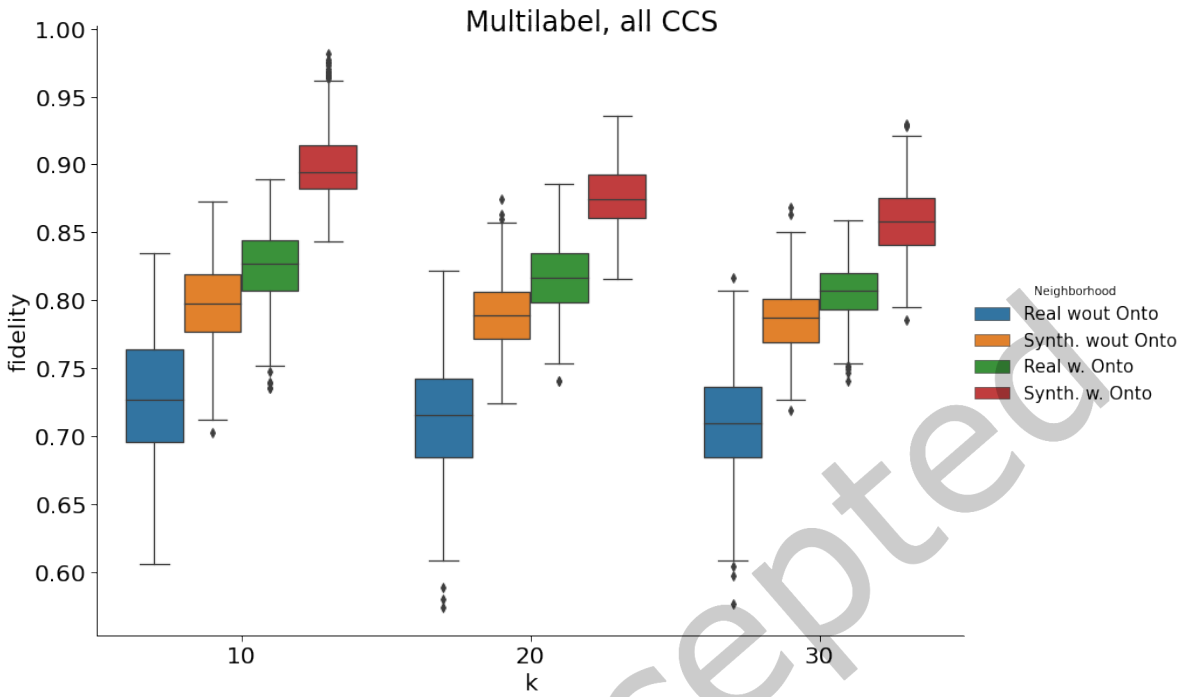


Fig. 6. Fidelity distribution at $k = 10, 20, 30$ for the explanation of the multi-label classification task. Each color represent a different condition.

boundary of the black box when we include in its training set patients that are distant from the patient under analysis, i.e. when trying to approximate a decision boundary that is not *local* anymore. The mean values of the *hit* metric at $k = 10, 20, 30$ for the multi-label classification task with and without the ontological information are all *hit*$= 1$, indicating that each condition allows the explainer to correctly classify the patient under analysis. Finally, given the high values of fidelity of the explanation pipeline that exploits the ontological information (red boxplot of figure 6), a one-way ANOVA was performed to compare the effect of k on *rule complexity* for this explanation pipeline. This test revealed that there was not a statistically significant difference in rule complexity between at least two groups ($F = 2.028$, $p = 0.131$).

Binary classification task. For the binary explanation task we selected all the patients predicted by the black box as having an *acute mycordial infarction* (CCS 100, 33 patients) *and pulmonary heart disease* (CCS 103, 121 patients). Given the higher values of fidelity for $k = 10$ of the multi-label classification task, we set $k = 10$ for all the experiments of the binary classification task increasing it only in the case in which all the first k real neighbors had the same classification. Similarly to the multi-label classification task, we show in figure 7 the fidelity distribution under different conditions: for the decision trees trained directly on the first k neighbors selected with the Jaccard (blue boxplot) and ontological (greed boxplot) distance, and the decision trees trained on synthetic neighbors generated with the normal (yellow boxplot) and ontological perturbation (red boxplot). We can observe the same trends observed for the multi-label classification task, confirming that in both cases,

Neighborhood	k	fidelity avg±std	k groups	mean diff.	p-adj	95% C.I. lower	95% C.I. upper
Real wout Onto	10	0.73 ± 0.05	10 and 20	-0.016	0.001(***)	-0.022	-0.010
	20	0.71 ± 0.04	10 and 30	-0.020	0.001(***)	-0.026	-0.014
	30	0.71 ± 0.04	20 and 30	-0.004	0.352	-0.001	0.003
Synth. wout Onto	10	0.80 ± 0.03	10 and 20	-0.008	0.001(***)	-0.012	-0.004
	20	0.79 ± 0.03	10 and 30	-0.011	0.001(***)	-0.015	-0.007
	30	0.77 ± 0.02	20 and 30	-0.003	0.322	-0.007	0.002
Real w. Onto	10	0.83 ± 0.03	10 and 20	-0.010	0.001(***)	-0.013	-0.006
	20	0.82 ± 0.02	10 and 30	-0.019	0.001(***)	-0.022	-0.015
	30	0.81 ± 0.02	20 and 30	-0.009	0.001(***)	-0.013	-0.006
Synth. w. Onto	10	**0.90 ± 0.03**	10 and 20	-0.023	0.001(***)	-0.027	-0.020
	20	**0.88 ± 0.02**	10 and 30	-0.041	0.001(***)	-0.044	-0.037
	30	**0.86 ± 0.03**	20 and 30	-0.017	0.001(***)	-0.021	-0.013

Table 2. Multi-label classification task. (*Left side of the table*) Mean and standard deviation of the fidelity for each neighborhood at different values of k. In bold the highest performance for every value of k. (*Right side of the table*) Results from the Tukey's HSD test for multiple comparisons between the means at different values of k for each neighborhood.

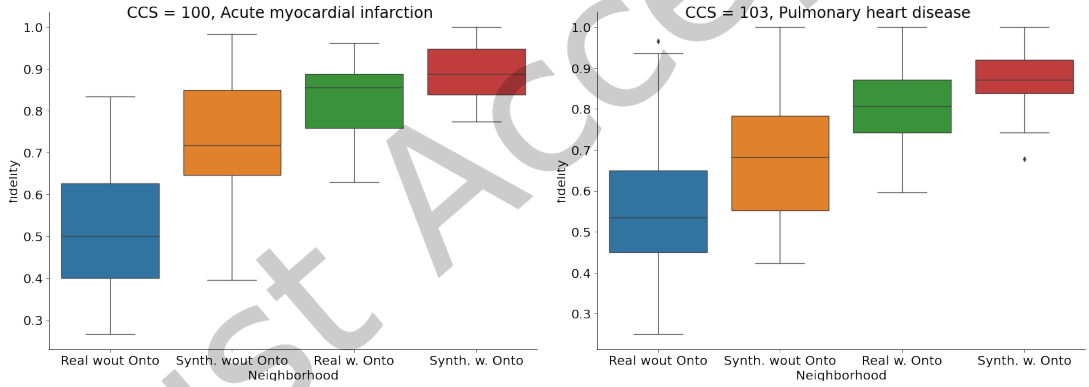


Fig. 7. Fidelity distribution for the explanation of the binary classification task. Each color represent a different condition. *(1st plot)* results for the CCS code 100, *acute myocardial infarction. (2nd plot)* results for the CCS code 103, *pulmonary heart disease.*

increasing the neighborhood density through perturbation and exploiting the ontology in the explanation process increase the explainer fidelity to the black box.

5 DOCTOR XAI EXPLANATIONS

In this section, we present the explanation provided by Doctor XAI. As described in section 3, Doctor XAI extracts a rule-based explanation from the (binary or multi-label) decision tree trained on the synthetic neighborhood. This rule is the decision path taken for the patient of interest on the decision tree. More formally, the explanation of the black box prediction $b(x) = y$ for the patient x is a decision-rule $p \rightarrow y$ whose premise p is the conjunction

of all the split conditions on the path from the root to the leaf node that is satisfied by the instance x on the decision tree. These split conditions are inequalities that follow the pattern:

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

Consider the following as an example of an explanation for the binary outcome *CCS-100: acute myocardial infarction* of one particular patient of interest:

$$\begin{aligned} p = &\{\text{ICD-9\_}V45.82 = 0.75 > 0.3125, \\ &\text{ICD-9\_}414.01 = 0.75 > 0.375, \\ &\text{ICD-9\_}357.2 = 0 \leq 0.34375\} \rightarrow y = \{\text{CCS-100}\} \end{aligned} \tag{1}$$

As you can see, both the *observed_value*s (0.75, 0.75 and 0) and the *threshold_value*s (0.3125, 0.375 and 0.34375) range between 0 and 1 and follow the temporal encoding representation as described in Section 3.5. Such representation assigns each code an exponentially decaying relevance (or *signal*) according to the visits they appear in: a signal of 0.5 is assigned to the codes appearing in the last visit, 0.25 to the second-to-last, and so on. With this logic, an inequality stating that ICD-9_code = $observed\_value < 0.25$ can be easily interpreted as *that ICD-9 code was not diagnosed in the last two visits*. However, such a raw rule-based explanation is hardly interpretable, we therefore need to process it to make it more human-understandable. Next sections present the translation in natural language (section 5.1) and the final user interface (section 5.2).

5.1 Natural Language Explanations

The first step towards a comprehensible output is performed by the rewriting of Doctor XAI's inequality-based explanations into natural language sentences. The algorithm for natural language rewriting encompasses the steps of *threshold adjustment*, *redundancy resolution*, and *string mapping*. The first step, *threshold adjustment*, is necessary as the *threshold_values* are produced by the split conditions of the decision tree, and they might not be easily interpretable. If the produced inequality is of the *less-than* kind, e.g.

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

We rewrite it as

$$\text{ICD-9\_code} = observed\_value < \overline{threshold\_value}$$

Where

$$\overline{threshold\_value} = min(\{x \mid x \in \{0, .25, .5, .75, 1.\} \wedge x > threshold\_value\})$$

Essentially, we weaken the inequality with the closest threshold that can be easily interpretable. Similarly, if the produced inequality is of the *greater-than* kind, e.g.

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

We rewrite it as

$$\text{ICD-9\_code} = observed\_value \geq \overline{threshold\_value}$$

Where

$$\overline{threshold\_value} = max(\{x \mid x \in \{0, .25, .5, .75, 1.\} \wedge x < threshold\_value\})$$

This step allows us to reduce the thresholds to a fixed sets, so that we can then encode them in natural language. Since some decision trees can produce paths with multiple split conditions over the same attribute, we can have multiple inequalities over the same code, e.g.:

$$(\text{ICD-9\_code} = observed\_value < t_1) \wedge (\text{ICD-9\_code} = observed\_value < t_2)$$

and therefore a *redundancy resolution* step is required. In this case we simply keep the strongest inequality, as the closest to the observed_value and therefore the most descriptive. So for the *less-than* kind of inequality we keep

the minimum t_i threshold, and for the *greater-than* we keep the maximum. Finally, we map our inequalities to natural language in the following way:

- observed_value < .25 : *was not diagnosed in the last two visits*
- observed_value < .5 : *was not diagnosed in the last visit*
- observed_value < .75 : *was not consistently diagnosed in the last visits*
- observed_value < 1. : *was not consistently diagnosed for enough visits*
- observed_value >= 0 : *was diagnosed at least once*
- observed_value >= .25 : *was diagnosed at least once in the last two visits*
- observed_value >= .5 : *was diagnosed in the last visit*
- observed_value >= .75 : *was consistently diagnosed in the last visits*

As an example, the whole process allows to transform the inequality-based premise of explanation (1) into the following natural language explanation:

```
Percutaneous transluminal coronary angioplasty status (V45.82) was diagnosed at least once in the
    last two visits, AND Coronary atherosclerosis of native coronary artery (414.01) was diagnosed
    at least once in the last two visits, AND Polyneuropathy in diabetes (357.2) was not diagnosed
    in the last visit THEREFORE Acute Myocardial Infarction (100).
```

5.2 Explanation user interface

This section describes the first prototype of the user interface created for Doctor XAI explanations. This interface lets the user explore the AI model behaviour through a dynamic visualization of the explanation. As described in section 4.1, the AI model is a clinical DSS that predicts all the conditions (ICD-9) that will be diagnosed in the following visit of the patient. The explainer is able to provide for both multi-label prediction tasks and for binary tasks. In this context, the multi-label prediction task involves predicting all the future diagnoses of a patient. The explanation interface allows the user to choose between the explanation of the multi-label or binary outcome. The ability to single out one condition and obtain an explanation for it is especially important when such a condition is particularly concerning, for example, if the model predicts that the patient will have an acute myocardial infarction. We show a static visualization of the explanation for a binary outcome (*acute myocardial infarction*) in figure 8.

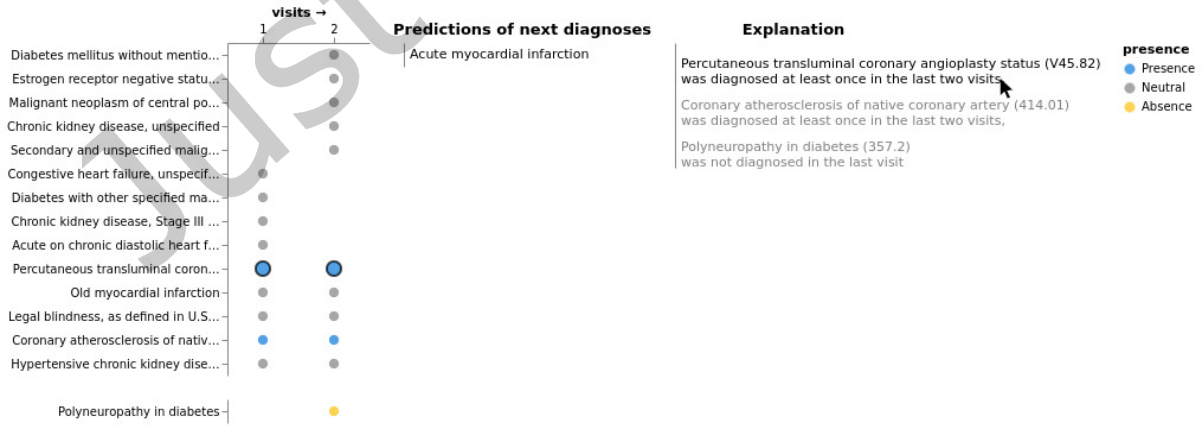


Fig. 8. Doctor XAI interface for binary prediction tasks

The explanation is represented as a chart where each condition (ICD-9) diagnosed in a visit is represented with a visual mark: a circle. All diagnoses are mapped on the vertical axis. The sequence of visits is mapped on the horizontal axis. Each circle allows the user to easily identify when each diagnosis was diagnosed (i.e. to which columns it belongs) and if it was encountered multiple times (i.e. multiple circles on the same row). To improve the identification of recurrent diagnoses, the conditions are sorted by their frequency, i.e. how many visits they appeared in. In this example, the patient was diagnosed with nine conditions in his first visit and ten conditions in the second.

Doctor XAI assigns a color to the grey dots representing the conditions according to their relevance for the black box prediction: dots corresponding to conditions deemed irrelevant are left grey, while dots deemed relevant are colored blue. Finally, Doctor XAI shows as yellow dots conditions that are missing from the patient's clinical history that would have changed algorithmic suggestion. The color scheme of the dots is colorblind-safe. The explanation expressed in natural language (see section 5.1) is shown after the algorithmic suggestion (*predictions of next diagnoses*), which is shown on the right side of the clinical history. The interactive interface allows the user to explore the textual explanation linked with the visualization, highlighting the occurrences of diagnosed in each visit when hovering with the mouse over the textual description of the explanation. An example of this is show in figure 8 where the mouse cursor is on one sentence of the explanation (in bold), and the associated dots are highlighted by black circles. Without loss in generality, the visualization can accommodate the case of multi-prediction task, reporting the list of all predicted symptoms for the next event.

6 HUMAN VALIDATION OF THE EXPLAINER

This section presents the first human validation step of the iterative design process. In particular, in this first step of human validation, we limit our study to the binary classification task. We carried out an online user study to understand the impact of the explanations provided by the prototype explainer presented in section 3 on healthcare providers. The aim of this first user study is to explore the relationship between trust and AI explanations and provide insights to inform next steps of the iterative design process. In particular, we focused on the following research questions:

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

Testing the following hypotheses:

- **Hp1**: Participants implicitly trust more the algorithmic suggestion when 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 explicitly express a higher trust in the system that provides an explanation.

6.1 Methods

6.1.1 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.

6.1.2 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. 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) [120, 121]. 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 our use case the clinical DSS is the advisor and the clinician is the judge, solely responsible to provide appropriate care for the patient.

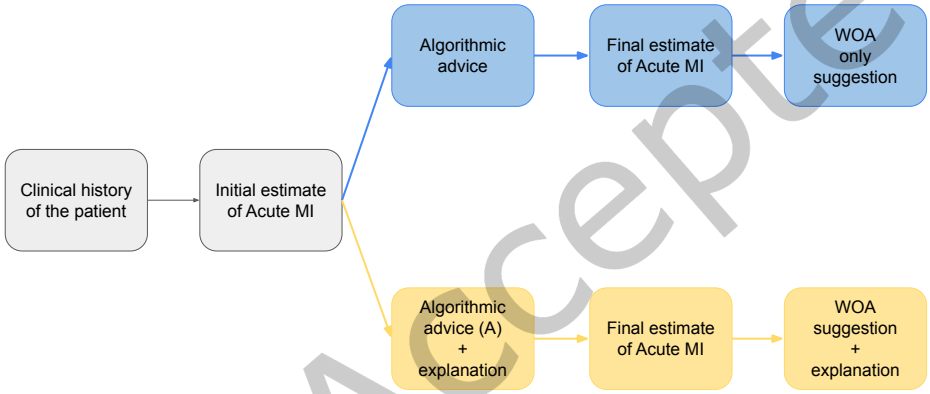


Fig. 9. Flowchart of the estimation task for the two interfaces: only suggestion (blue path) and suggestion and explanation (yellow path)

6.1.3 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 9) and once using the interface providing the suggestion and the explanation (yellow path of figure 9). 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).

6.1.4 Collected data.

Implicit trust and confidence. In the context of decision-making, trust is positively associated with advice taking [53, 121]. Our main dependent variable is therefore the Weight of Advice (WOA) [60] 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") [3, 25, 64].

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 [132, 133]. 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 [133].
- **Effort Expectancy**: the degree of ease associated with the use of the system [133].
- **Attitude Towards using Technology**: an individual's overall affective reaction to using a system [133].
- **Social Influence**: the degree to which an individual perceives that important others believe he or she should use the new system [133].
- **Facilitating Conditions**: the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system [133].
- **Image**: the degree to which use of an innovation is perceived to enhance one's image or status in one's social system [132].
- **Job relevance**: The degree to which an individual believes that the target system is applicable to his or her job [132].
- **Output Quality**: The degree to which an individual believes that the system performs his or her job tasks well [132].
- **Result Demonstrability**: The degree to which an individual believes that the results of using a system are tangible, observable, and communicable [132].

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 [64] and collected qualitative feedback using open-ended question on participants' experience using the two AI interfaces.

Confounding factors. We controlled for confounding factors such as participants' familiarity and involvement in the task [44], demographic information such as gender, age, and the type of medical profession. 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") [24, 90].

6.2 Results

6.2.1 Quantitative analysis. A total of 45 healthcare providers participated in the online experiment. The analysis discarded three participants: one did not pass the attention check question, while three gave 100 as their initial estimate, which yielded undefined values for the WOA ($A = I$). Eventually, 41 participants were retained for the study. 11 doctors, 26 nurses, one health care assistant, one dietetic assistant practitioner, one ambulance call dispatcher, and one paramedic. The mean age was 39 years old (SD=12) ranging from 24 to 65 years old. 31 were

women and 10 men. The male sample had a mean age of 32 years old (SD=8), and the female sample had mean 42 years old (SD=12). We performed all the analysis in Python.

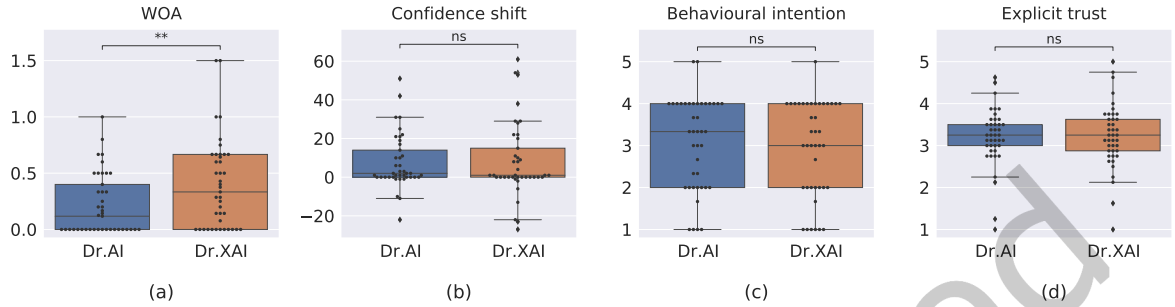


Fig. 10. Boxplot comparing the WOA (a) the confidence shift after the advice (b) the behavioural intention of use and (c) the explicit trust in the two systems (d).

Weight of Advice and Confidence. . In figure 10(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=0.33) than the Dr.AI interface (Mdn=0.12). A paired-samples two-sided Wilcoxon signed-rank test indicated that this difference was statistically significant ($T = 103.5$, $p = 0.003$). 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 10(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 $T = 313.5$, $p = 0.566$. 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 10(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 $T = 99$, $p = 0.231$. 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 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 $rs(27) = 0.60$, $p < .001$. Similarly, we did not find a significant difference in explicit trust between the two interfaces (figure 10(d), paired-samples two-sided Wilcoxon signed-rank test, $T = 303.0$, $p = 0.461$), but we found a strong positive Spearman correlation between explicit trust and perceived explanation quality ($rs(27) = 0.72$, $p < .001$). 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. As regards further results regarding the questionnaires, see Appendix.

6.3 Open-ended questions insights

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 and 4) How would they change the explanation user interface. Almost all users preferred the AI interface that provided the explanation for its outcome. When asked what they liked about it, many mentioned they liked the fact that more information was provided compared to the interface without an explanation:

"[I like that it provides] *some detail of what it considers important in the calculation of risk.*"

"[I like] *that it gave an explanation of sorts*"

"*A lot more information is available.*"

However, when asked what they disliked more about the explanation interface, many participants mentioned it contained too much information and not necessarily the information they wanted:

"*Slight information overload and made the decision making much slower than usual.*"

"*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.*

"*In theory I like having more info but i appear not to have grasped it.*"

When asked how would they improve the explanation, many mentioned that they found the ICD-9 codes confusing and that they would have preferred a more textual explanation:

"*the ICD codes are confusing.*"

[I would like] "*less use of icd codes*"

"*Give an explanation in human language why the program thinks an MI is likely and HOW likely in %*"

"*make it more textual. it is not an explanation, it just highlights the higher risk codes in the patients' record.*"

"*would like to see an explanation like, 'This patient is X% likely to undergo an acute MI within Y time. Suggest steps A,B and C'*"

Many also suggested simplifying the interface:

"*List only those codes that are pertinent to a cardiovascular diagnosis*"

"*Make it simple*"

"*simplify*"

"[I would like]*a clearer and more precise explanation*"

While others felt they did not have all the relevant clinical information they needed:

"*I would include the length of time between visits to allow a clinician to more accurately interpret the results*"

"[I would improve it by] *Using investigation outcomes, symptoms, patient data such as age/weight/family history to give a more holistic outcome*"

Finally, many wanted information on the reliability of the suggestion:

"[I don't like] the uncertainty around accuracy "

"We don't know how reliable it is yet as we haven't got any success rate information"

"It could be wrong, and could lead to mistakes or things being missed"

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. 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).

6.3.1 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, 3, 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).
>
> *Probability outcomes are useful in diagnosis but so is our human ability to reason and to understand that humanity, free will and determination to overcome surprises us constantly which is something no AI can understand nor predict.* (Dr.XAI condition)

While this might be associated with the phenomenon of *algorithm aversion* [40], or the human discount of algorithmic advice [85], 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 [46].

7 REDESIGN OF THE EXPLANATION USER INTERFACE

This section presents the redesign of the explanation user interface based on users feedback received from the open questions of the human validation step. The new interface is designed as follow. The doctor is initially shown a home screen where (s)he can visualize all the patients under his/her care (figure 11 (a)). Each patient is represented by a card containing their name and their risk of having an acute myocardial infarction (MI). Patients classified by the AI system as having a high risk are highlighted to catch the doctor's attention (in yellow in the interface).

If the doctor wants to have more information on one particular patient, (s)he can click on the "Discover more" button and visualize the basic demographics of the patient (figure 11 (b)). At this step (s)he can also ask to have an explanation for the AI system prediction by clicking on the "Get the explanation" button. The explanation is first presented in natural language, i.e., using a sentence that uses the names of the relevant diseases and that mentions when they were diagnosed, without displaying the ICD codes, e.g. 413.9, (figure 11 (c)). However, if the doctor wants, (s)he can visualize the diagnoses considered relevant for the AI system prediction as a temporal sequence of events (the colored dots of the initial interface, figure 11 (d)) by clicking on the "Show symptoms" button. Finally, the doctor can decide to visualize the full clinical history of the patient and explore each clinical event with the mouse-over (the complete initial interface having both colored and gay dots, figure 11 (e)). Only at this point, the ICD-9 codes are shown to the doctor when s(he) drag the mouse over the dots representing the diagnosis. This new interactive explanation user interface can be explored on the website https://kdd.isti.cnr.it/DrXAI-viz/.

The key feature of this new interface is the *progressive disclosure* of information [34, 92, 123]. The initial interface provided all the relevant information to the user at once. This resulted in an overwhelming amount of information and a cluttered interface. Thanks to the progressive disclosure of information, the interface guides

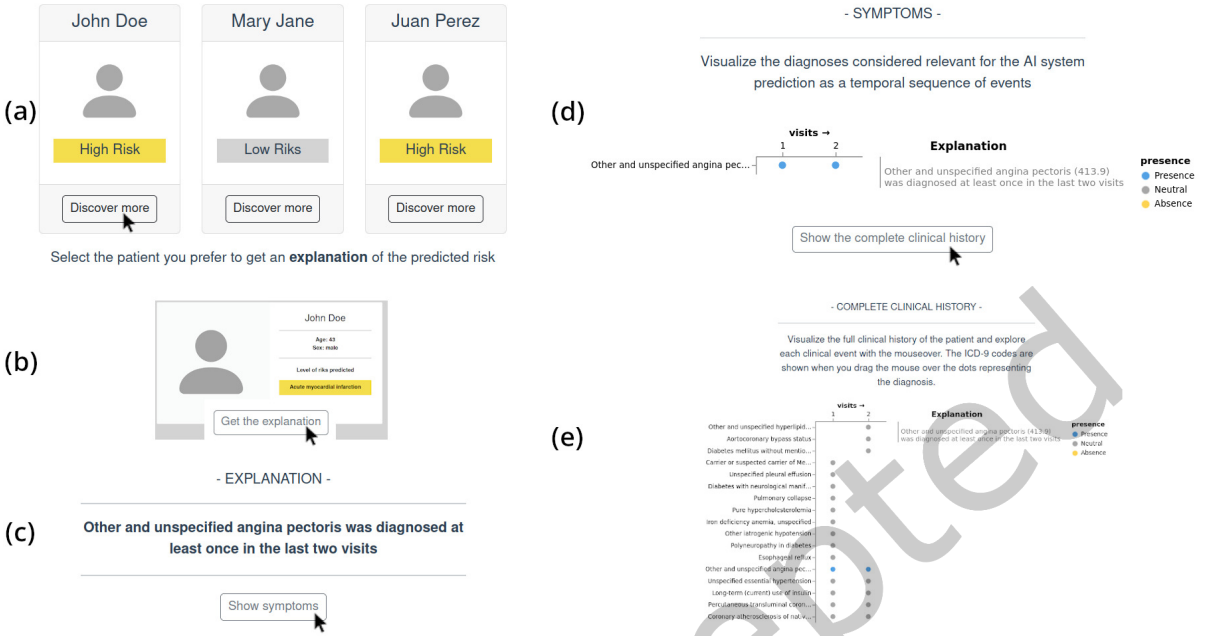


Fig. 11. New explanation user interface

doctors to explore the relevant factors in the AI decision to a degree they feel comfortable with. The explanation is provided only when asked, and the first type of explanation provided is the one in natural language. However, some participants highlighted the need to have more information on the patient. Therefore we left the option of visualizing the full clinical history of the patient as the last step of the explanation. The initial interface was prototyped using a technocentric approach to AI explanations. Since the AI algorithm processed ICD codes to predict future conditions, the explanation was provided in terms of ICD codes. Using a human-centered approach to XAI, we decided to leave the ICD codes to the last step of the explanation, visualizing them only if the user explicitly hovers the mouse over the corresponding dots.

7.1 Heuristic Evaluation

We conducted an assessment of the two explanation interfaces through a *heuristic evaluation* [98], a usability testing method where a group of evaluators judges an interface using a set of usability guidelines. The primary goal was to compare the designs of the old and new interfaces. Ten members of our team participated in the heuristic evaluation. Participants were asked to assess the two interfaces according to Nielsen's ten usability heuristics: *visibility of the system status*; *match between system and the real world*; *user control and freedom*; *consistency and standards*; *error prevention*; *recognition rather than recall*; *flexibility*; *aesthetic and minimalist design*; *help for users to recognize and recover from errors*; and *help and documentation*. Both interfaces are at an early stage of prototyping, hence not all the 10 Nielsen's heuristics were evaluated. In particular, the following heuristics were not considered: 1) error prevention, 2) help users recognize and recover from errors, and 3) help and documentation. The reason behind this choice is due to the limited functionality of the interface that does not permit errors, and lacks a documentation. We asked the evaluators to check the different patients in

both interfaces to get insights into their comprehension of them. During the evaluation of the new interface, we received both positive and negative comments that we discuss below.

The new explanation interface implemented following users' feedback from the open-ended questions has been evaluated more positively compared to the old one for the following heuristics:

- Visibility of the system status. All pages have been labeled clearly with the headers. One evaluator required a more visible button to go back to the upper part of the page.
- Match between the system and the real world. All the icons were clear. One evaluator found the terminology too technical but this can be expected since the evaluators did not have a medical background.
- User control and freedom. Since the functionality of the system is limited, the freedom and the control of the users are limited too.
- Consistency and standards. This heuristic has not been violated. The second interface has been evaluated as clearer and more organized in the information disclosure.
- Recognition rather than recall. The second interface has been evaluated as easier to use because of the progressive disclosure of the patients' information. This is expected since progressive disclosure can reduce mental workload.
- Aesthetics and minimalist design. This heuristic was evaluated positively for both the interfaces because the aesthetics has not changed.

During the assessment, the negative aspects were related to:

- Flexibility. As reported before in the visibility heuristic, one evaluator found it difficult to see the button to go back to the upper part of the screen. The button can be difficult to use for inexperienced users.

The heuristics listed below, although not considered in the analysis because of the limited functionality of the current interface, were still mentioned by the evaluators as important points to focus on for future improvements of the interface.

- Help and documentation. The documentation is still lacking since the AI application is in its prototype stage, furthermore we relied on conventions to design the visual interface.
- Error prevention. We did not consider this heuristics because the interface offers a limited set of actions and does not allow the users to make errors.
- Help the users recognize and recover from errors. As described above, the interface we tested gives not the opportunity to make errors, hence this heuristic has not been considered.

In conclusion, the heuristic evaluation showed that the second interface is considered more usable compared to the first one, even if a few pointers were raised to improve the flexibility of the system and the visibility of some elements of the interface.

8 CONCLUSIONS

Co-designing human-centered AI explanations involves working with end users to create explanations that enable them "*to achieve goals effectively, efficiently and with satisfaction, taking account of the context of use*" [65]. This paper showcase this process presenting a cycle of prototyping, testing, and redesigning of a XAI method and its user interface for clinical Decision Support Systems (DSSs). The technical development of the XAI method, which is designed to handle technical features commonly found in healthcare settings, such as sequential and ontology-linked data and multi-label tasks, is described in Section 3. The technical requirements of the system are then evaluated in Section 4, and the creation of a first user interface for its explanations is illustrated in Section 5. Trust is a crucial factor in the acceptance and use of clinical DSS, and AI explanations can play a central role in building trust by allowing healthcare providers to inspect the factors that led the AI system to make a particular recommendation. However, AI explanations can also increase over-reliance on the system

[47, 136]. To investigate the relationship between AI explanations and trust in the system, as well as participants' perceptions of the interface, we conducted a user study in which we gathered feedback and observations from healthcare providers in Section 6. This feedback was then used to co-design a more human-centered explanation interface in Section 7. Finally, we performed a heuristic evaluation to compare the usability of the two interfaces in Section 7.1.

The user study found that participants were more likely to follow the AI system's advice when it explained its suggestion, as reflected in the greater shift in participants' estimates measured by the Weight Of Advice (WOA) (figure 10(a)). This suggests that explanations may help to increase trust in the AI system. However, the study only considered correct suggestions from the AI system. Hence, it is unclear whether explanations would have the same effect on trust when the AI system gives incorrect suggestions. In other words, more research is needed to understand if AI explanations can effectively help in the trust calibration process. It is worth noting that participants showed an increase in implicit trust in the system when presented with an AI explanation despite experiencing information overload and frustration with the initial explanation interface (section 6.3) raising concerns about the potential for automation bias. The user study also investigated confidence after receiving advice and explicit trust in the system but did not find any significant difference between the two AI interfaces (with and without explanations) (figure 10(b)). The study also found no significant difference in behavioral intention (BI) (figure 10(c)). One possible explanation for this might lie in the high correlation we measured between BI and explanation quality (section 6.2.1), indicating that the proposed explanations were probably ill-suited for a healthcare audience. It is also important to note that the WOA was the only implicit measure in the study and was the only measure that showed a statistically significant difference between the two interfaces (figure 10). All other measures, including confidence and BI, were explicit measures that captured intentional behaviors, which may suggest that participants were unaware of the influence of the explanations on their estimates. The results of the heuristic evaluation showed that the new explanation interface, which was redesigned based on feedback received in open-ended questions, was more usable compared to the first interface (section 7.1). However, a more detailed assessment involving healthcare professionals may be needed to determine the impact of the new interface on trust and decision performance. These findings have implications for the design of explanation interfaces, highlighting the importance of considering user feedback and the potential benefits of testing interfaces with target users. Participants also expressed a need to understand the uncertainty associated with the AI system's advice (section 6.3). However, this feedback was not implemented in the new AI explanation interface because it was a limitation of the underlying AI system and XAI technique, rather than the interface itself. In an ideal scenario, all clinical DSS should clearly communicate the uncertainty around their recommendations. Unfortunately, the AI system used in our proof of concept did not provide this information. Furthermore, it is important to distinguish between explaining AI outcomes and explaining the uncertainty associated with them. This insights highlight the importance of an iterative design process, as it can uncover new technical requirements that can be addressed in future iterations.

ACKNOWLEDGMENTS

This work was partially supported by the European Union under ERC-2018-ADG G.A. 834756 (XAI), G.A. 952026 (HumanE AI Net), PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI". This work was also partially supported by Horizon Europe project G.A. 101057746 (PRE-ACT), by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22 00058, and by the UK government (Innovate UK application number 10061955).

REFERENCES

[1] European Commission 2018. *EU General Data Protection Regulation.* European Commission. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[2] 2021. *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.* https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[3] Barbara D Adams, Lora E Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann. 2003. Trust in automated systems. *Ministry of National Defence* (2003).

[4] Bibb Allen, Sheela Agarwal, Laura Coombs, Christoph Wald, and Keith Dreyer. 2021. 2020 ACR Data Science Institute Artificial Intelligence Survey. *Journal of the American College of Radiology* (2021).

[5] Omar AlShorman, Buthaynah Alshorman, and Fahed Alkahtani. 2021. A review of wearable sensors based monitoring with daily physical activity to manage type 2 diabetes. *International Journal of Electrical and Computer Engineering (IJECE)* 11, 1 (2021), 646–653.

[6] Ahmad Fayez S Althobaiti. 2017. Comparison of Ontology-Based Semantic-Similarity Measures in the Biomedical Text. *Journal of Computer and Communications* 5, 02 (2017), 17.

[7] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences* 11, 11 (2021), 5088.

[8] 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. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[9] Robert Avram, Jeffrey E Olgin, Peter Kuhar, J Weston Hughes, Gregory M Marcus, Mark J Pletcher, Kirstin Aschbacher, and Geoffrey H Tison. 2020. A digital biomarker of diabetes from smartphone-based vascular signals. *Nature Medicine* 26, 10 (2020), 1576–1582.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *CoRR* abs/2006.14779 (2020). arXiv:2006.14779 https://arxiv.org/abs/2006.14779

[11] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence* (2021), 1–10.

[12] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics* 22, 7 (2006), 830–836.

[13] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

[14] Donald J. Berndt and James Clifford. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (Seattle, WA) *(AAAIWS'94)*. AAAI Press, 359–370. http://dl.acm.org/citation.cfm?id=3000850.3000887

[15] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[16] Alan F. Blackwell. 2021. Ethnographic artificial intelligence. *Interdisciplinary Science Reviews* 46, 1-2 (2021), 198–211. https://doi.org/10.1080/03080188.2020.1840226

[17] 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. 2018. 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 (2018), 1160–1166.

[18] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004).

[19] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076* (2021).

[20] Andrea Brennen. 2020. 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*. 1–7.

[21] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[22] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. 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 (2021), 1–21.

[23] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[24] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.

[25] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety science* 47, 9 (2009), 1260–1270.
[26] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " 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 (2019), 1–24.
[27] Giacomo Cappon, Martina Vettoretti, Giovanni Sparacino, and Andrea Facchinetti. 2019. Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications. *Diabetes & metabolism journal* 43, 4 (2019), 383–397.
[28] Donna J Cartwright. 2013. Icd-9-cm to icd-10-cm codes: what? why? how?
[29] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789
[30] Benjamin Chin-Yee and Ross Upshur. 2020. The Impact of Artificial Intelligence on Clinical Judgment: A Briefing Document. (2020).
[31] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*. PMLR, 301–318.
[32] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 787–795.
[33] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. 2020. Multi-label transfer learning for the early diagnosis of breast cancer. *Neurocomputing* 392 (2020), 168–180.
[34] Michael Chromik and Andreas Butz. 2021. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *IFIP Conference on Human-Computer Interaction*. Springer, 619–640.
[35] Amanda Clare and Ross D King. 2001. Knowledge discovery in multi-label phenotype data. In *European conference on principles of data mining and knowledge discovery*. Springer, 42–53.
[36] Giovanni Comandé. 2020. Unfolding the legal component of trustworthy AI: a must to avoid ethics washing. *Version Accepted for Annuario di Diritto Comparato e di Studi Legislativi, forthcoming* (2020).
[37] Ian Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research* 22, 209 (2021), 1–90. http://jmlr.org/papers/v22/20-1316.html
[38] Ian Covert, Scott M Lundberg, and Su-In Lee. 2021. Explaining by Removing: A Unified Framework for Model Explanation. *J. Mach. Learn. Res.* 22 (2021), 209–1.
[39] Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems* 8 (1995), 24–30.
[40] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.
[41] Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics* 116 (2021), 103728.
[42] Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.
[43] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[44] Jinyun Duan, Yue Xu, and Lyn M Van Swol. 2020. Influence of self-concept clarity on advice seeking and utilisation. *Asian Journal of Social Psychology* (2020).
[45] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[46] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
[47] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
[48] Shaker El-Sappagh and Farman Ali. 2016. DDO: a diabetes mellitus diagnosis ontology. In *Applied Informatics*, Vol. 3. Springer, 5.
[49] Wenjuan Fan, Jingnan Liu, Shuwan Zhu, and Panos M Pardalos. 2018. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research* (2018), 1–26.
[50] Ruiwei Feng, Yan Cao, Xuechen Liu, Tingting Chen, Jintai Chen, Danny Z Chen, Honghao Gao, and Jian Wu. 2021. ChroNet: A multi-task learning based approach for prediction of multiple chronic diseases. *Multimedia Tools and Applications* (2021), 1–15.
[51] Ruiquan Ge, Renfeng Zhang, and Pu Wang. 2020. Prediction of Chronic Diseases With Multi-Label Neural Network. *IEEE Access* 8 (2020), 138210–138216.

[52] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[53] Francesca Gino and Maurice E Schweitzer. 2008. Take this advice and shove it. In *Academy of Management Proceedings*, Vol. 2008. Academy of Management Briarcliff Manor, NY 10510, 1–5.

[54] Dominic Girardi, Sandra Wartner, Gerhard Halmerbauer, Margit Ehrenmüller, Hilda Kosorus, and Stephan Dreiseitl. 2016. Using concept hierarchies to improve calculation of patient similarity. *Journal of biomedical informatics* 63 (2016).

[55] 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. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.

[56] Thomas R Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition* 5, 2 (1993), 199–220.

[57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models. *ACM CSUR* 51, 5, Article 93 (Aug. 2018), 42 pages.

[58] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 549–559.

[59] Yukinori Harada, Shinichi Katsukura, Ren Kawamura, and Taro Shimizu. 2021. 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 (2021), 5562.

[60] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133.

[61] John P Higgins. 2016. Smartphone applications for patients' health and fitness. *The American journal of medicine* 129, 1 (2016), 11–19.

[62] HIMSS. 2019. *AI use in European healthcare - HIMSS*. https://www.himssanalytics.org/europe/ehealth-barometer/ehealth-trend-barometer-ai-use-european-healthcare [Accessed: 2021-09-20].

[63] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[64] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[65] ISO 9241-210:2019 2019. *Ergonomics of human-system interaction Human-centred design for interactive systems*. Standard. International Organization for Standardization.

[66] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[67] Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13, 6 (2012), 395–405.

[68] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. 2009. Use of electronic health records in US hospitals. *New England Journal of Medicine* 360, 16 (2009), 1628–1638.

[69] Zheng Jia, Xudong Lu, Huilong Duan, and Haomin Li. 2019. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 91. https://doi.org/10.1186/s12911-019-0807-y

[70] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Mark Roger. 2020. MIMIC-IV (version 0.4). *PhysioNet* (2020). https://doi.org/10.13026/a3wn-hq05

[71] 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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.

[72] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

[73] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[74] Ajay Kohli and Saurabh Jha. 2018. Why CAD failed in mammography. *Journal of the American College of Radiology* 15, 3 (2018), 535–537.

[75] Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harmander Kaur. 2018. The use of electronic health records to support population health: a systematic review of the literature. *Journal of medical systems* 42, 11 (2018), 1–16.

[76] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[77] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine* 94 (2019), 42–53.
[78] Frank Lawler, Jim R Cacy, Nancy Viviani, Robert M Hamm, and Stephen W Cobb. 1996. Implementation and termination of a computerized medical information system. *Journal of Family Practice* 42, 3 (1996), 233–236.
[79] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[80] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.
[81] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. 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*. 1–13.
[82] Runzhi Li, Wei Liu, Yusong Lin, Hongling Zhao, and Chaoyang Zhang. 2017. An ensemble multilabel classification for disease risk prediction. *Journal of healthcare engineering* 2017 (2017).
[83] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[84] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[85] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[86] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.
[87] Gianclaudio Malgieri and Giovanni Comandé. 2017. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law* (2017).
[88] Vidushi Marda and Shivangi Narayan. 2021. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence* 3, 3 (2021), 187–189.
[89] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. 2021. Exemplars and Counterexemplars Explanations for Image Classifiers, Targeting Skin Lesion Labeling. In *2021 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–7.
[90] Martijn Millecamp, Sidra Naveed, Katrien Verbert, and Jürgen Ziegler. 2019. 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*, Vol. 2450. CEUR; http://ceur-ws. org/Vol-2450/paper2. pdf, 10–18.
[91] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 417–431.
[92] Johanna D Moore and Cécile L Paris. 1991. Requirements for an expert system explanation facility. *Computational Intelligence* 7, 4 (1991), 367–370.
[93] Jessica Morley, Caio CV Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine* (2020), 113172.
[94] Annette Moxey, Jane Robertson, David Newby, Isla Hains, Margaret Williamson, and Sallie-Anne Pearson. 2010. Computerized clinical decision support for prescribing: provision does not guarantee uptake. *Journal of the American Medical Informatics Association* 17, 1 (2010), 25–33.
[95] Henrik Mucha, Sebastian Robert, Ruediger Breitschwerdt, and Michael Fellmann. 2021. 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*. 1–6.
[96] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.
[97] Emanuele Neri, Francesca Coppola, Vittorio Miele, Corrado Bibbolino, and Roberto Grassi. 2020. Artificial intelligence: Who is responsible for the diagnosis?
[98] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/97243.97281
[99] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. 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*, Vol. 8. 112–121.
[100] Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. 2019. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*. Springer, 97–110.
[101] Cecilia Panigutti, Anna Monreale, Giovanni Comandè, and Dino Pedreschi. 2022. Ethical, Societal and Legal Issues in Deep Learning for Healthcare. In *Deep Learning In Biology And Medicine*. World Scientific, 265–313.

[102] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 2021. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management* 58, 5 (2021), 102657.
[103] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 629–639.
[104] Jonathan M Peake, Graham Kerr, and John P Sullivan. 2018. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in physiology* 9 (2018), 743.
[105] Ben Joseph Philip, Mohamed Abdelrazek, Alessio Bonti, Scott Barnett, and John Grundy. 2022. Data Collection Mechanisms in Health and Wellness Apps: Review and Analysis. *JMIR mHealth and uHealth* 10, 3 (2022), e30468.
[106] Jennifer Preece, Yvonne Rogers, and Helen Sharp. 2019. *Interaction Design: Beyond Human-Computer Interaction* (5 ed.). Wiley.
[107] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics* 83 (2018), 112–134.
[108] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.
[109] Kathryn Rough, Andrew M Dai, Kun Zhang, Yuan Xue, Laura M Vardoulakis, Claire Cui, Atul J Butte, Michael D Howell, and Alvin Rajkomar. 2020. Predicting inpatient medication orders from electronic health record data. *Clinical Pharmacology & Therapeutics* 108, 1 (2020), 145–154.
[110] Murali Sambasivan, Pouyan Esmaeilzadeh, Naresh Kumar, and Hossein Nezakati. 2012. 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 (2012), 1–8.
[111] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 431–449.
[112] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* 40, D1 (2012), D940–D946.
[113] Jessica M Schwartz, Amanda J Moy, Sarah C Rossetti, Noémie Elhadad, and Kenrick D Cato. 2021. 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 (2021), 653–663.
[114] Ian A Scott, Ahmad Abdel-Hafez, Michael Barras, and Stephen Canaris. 2021. What is needed to mainstream artificial intelligence in health care? *Australian Health Review* (2021).
[115] Atul Sharma, Mihaela Badea, Swapnil Tiwari, and Jean Louis Marty. 2021. Wearable biosensors: an alternative and practical approach in healthcare and disease monitoring. *Molecules* 26, 3 (2021), 748.
[116] Lucy Shinners, Christina Aggar, Sandra Grace, and Stuart Smith. 2020. Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: an integrative review. *Health informatics journal* 26, 2 (2020), 1225–1236.
[117] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, et al. 2021. BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Medical Image Analysis* 71 (2021), 102046.
[118] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.
[119] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* 25, 11 (2007), 1251.
[120] Janet A Sniezek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62, 2 (1995), 159–174.
[121] Janet A Sniezek and Lyn M Van Swol. 2001. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes* 84, 2 (2001), 288–307.
[122] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. 2022. Metrics, Explainability and the European AI Act Proposal. *J* 5, 1 (2022), 126–138.
[123] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.
[124] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter PC Boon, and Ellen HM Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology* 30 (2020), 5525–5532.
[125] Rudi Studer, V Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: principles and methods. *Data & knowledge engineering* 25, 1-2 (1998), 161–197.
[126] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 1–10.

[127] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.
[128] Eric Topol. 2019. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.
[129] Madhukar H Trivedi, JK Kern, A Marcee, B Grannemann, B Kleiber, T Bettinger, KZ Altshuler, and A McClelland. 2002. Development and implementation of computerized clinical guidelines: barriers and solutions. *Methods of information in medicine* 41, 05 (2002), 435–442.
[130] Helena Varonen, Tiina Kortteisto, Minna Kaila, and EBMeDS Study Group. 2008. What may help or hinder the implementation of computerized decision support systems (CDSSs): a focus group study with physicians. *Family practice* 25, 3 (2008), 162–167.
[131] Viswanath Venkatesh. 2021. Adoption and use of AI tools: a research agenda grounded in UTAUT. *Annals of Operations Research* (2021), 1–12.
[132] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.
[133] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.
[134] Himanshu Verma, Roger Schaer, Julien Reichenbach, Jreige Mario, John O Prior, Florian Evéquoz, and Adrien Raphaël Depeursinge. 2021. On Improving Physicians' Trust in AI: Qualitative Inquiry with Imaging Experts in the Oncological Domain. (2021).
[135] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, 318–328.
[136] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[137] World Health Organization WHO et al. 2018. ICD Purpose and Uses. *Classification. Available online at: http://www.who. int/classifications/icd/en/(Accessed May 20, 2020)* (2018).
[138] Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 133–138.
[139] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. 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*. 1–13.
[140] Jiamin Yin, Kee Yuan Ngiam, and Hock Hai Teo. 2021. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. *Journal of medical Internet research* 23, 4 (2021), e25759.
[141] Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical Attention Propagation for Healthcare Representation Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 249–256.
[142] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering* 18, 10 (2006), 1338–1351.
[143] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2014).
[144] Xiaoqing Zhang, Hongling Zhao, Shuo Zhang, and Runzhi Li. 2019. A novel deep neural network model for multi-label chronic disease prediction. *Frontiers in genetics* 10 (2019), 351.
[145] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. 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*. 295–305.

APPENDIX

UTAUT variable	median Dr.AI	median Dr.XAI	Wilcoxon statistic	p-value
Performance Expectancy	3.2	3.2	184.5	0.913
Effort Expectancy	3.8	3.5	219.0	0.070
Social Influence	3.2	3.5	131.0	0.097
Facilitating Conditions	3.5	3.5	143.0	0.595
Attitude toward techology use	3.0	3.0	204.5	0.170
Image	2.0	2.0	78.5	0.504
Relevance	3.3	3.0	112.5	0.648
Output quality	3.0	3.0	137.5	0.134
Result Demonstrability	3.8	3.8	306.5	0.675
Behavioural intention	3.3	3.0	99.0	0.230

Table 3. Comparison of UTAUT variables for the two interfaces. Median, paired sample Wilcoxon signed-rank test statistics and p-value.

Further findings. In table 3 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. Following Bonferroni's correction method, a more stringent alpha (α=.005) was set for these particular tests. 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 ($rs(40)=-0.51$, 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 k-sample Anderson–Darling test showed a slight difference in the WOA for the Dr.AI interface between the different types of healthcare providers ($A = 1.986$ with significance level=0.047). However, given the small sample for each category, we leave such an analysis for further works.