CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis

Yao Xie
HCI Research, UCLA
Los Angeles, California,
United States
yaoxie@ucla.edu

Melody Chen
HCI Research, UCLA
Los Angeles, California,
United States
melodyc1205@ucla.edu

David Kao
HCI Research, UCLA
Los Angeles, California,
United States
davidkao41@ucla.edu

Ge Gao
College of Information
Studies, University of
Maryland, College Park
Maryland, United States
gegao@umd.edu

Xiang 'Anthony' Chen
HCI Research, UCLA
Los Angeles, California,
United States
xac@ucla.edu

ABSTRACT

The recent development of data-driven AI promises to auto-mate medical diagnosis; however, most AI functions as 'black boxes' to physicians with limited computational knowledge. Using medical imaging as a point of departure, we conducted three iterations of design activities to formulate CheXplain—a system that enables physicians to explore and understand AI-enabled chest X-ray analysis: *(i)* a paired survey between referring physicians and radiologists reveals whether, when, and what kinds of explanations are needed; *(ii)* a low-fidelity prototype co-designed with three physicians formulates eight key features; and *(iii)* a high-fidelity prototype evaluated by another six physicians provides detailed summative insights on how each feature enables the exploration and understand-ing of AI. We summarize by discussing recommendations for future work to design and implement explainable medical AI systems that encompass four recurring themes: motivation, constraint, explanation, and justification.

Author Keywords

Explainable artificial intelligence; physician-centered design; system design.

CCS Concepts

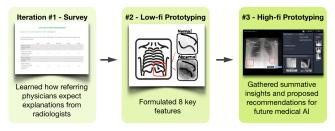•**Human-centered computing → Human computer inter-action (HCI)**;

Figure 1. We conducted three user-centered activities to formulate the design of CheXplain — a system that enables physicians to explore and understand AI-enabled chest X-ray analysis. *(i)* A paired survey study learns about how referring physicians expect explanations from radiolo-gists. *(ii)* A user-centered design with three physicians formulates eight design ideas as low-fi prototypes. *(iii)* An integrated high-fi prototype evaluated by six physicians provides insights and recommendations for future development of explainable medical AI systems.

INTRODUCTION

Intelligent systems—computational agents that employ artifi-cial intelligence (AI) to process and make sense of data—are becoming increasingly ubiquitous in modern workplaces [1]. For example, stakeholders use algorithms to assist them in urban planning, predicting the risk of future crimes or estimat-ing insurance risk [61, 69]. Despite the promise of assisting human decision making through a data-driven approach, non-computing professionals often find it challenging to under-stand how a 'black box' system transforms their initial input into a final decision.

In medicine, the 'black box' problem of AI also raised con-cerns. Started in the 1950s [44], medical AI has been widely applied in many subfields such as early detection, diagnosis, and treatment planning [35]. Early work on medical expert sys-tems such as DXplain [4] was able to incorporate rule-based explanation as part of the automatically generated results.

The recent development of deep neural networks promises that data-driven AI can process diagnostic imaging on par with hu-man physicians [32]. However, compared with expert systems,

neural networks rely on very low-level constructs—a large number of neurons distributed across stacks of layers, making it challenging, if not impossible, for medical professionals to understand how a data-driven AI arrives at certain conclusions. Such a lack of transparency creates a barrier of understanding, preventing AI from being widely adopted in the clinic despite of its promising performance [70].

To solve the 'black box' problem, prior work has been focusing on developing interpretable, accountable and transparent algorithms [16, 56], visualizing obscure features in medical images [17, 68] or employing cognitive psychological theories to explore effective explanations [41, 42, 44]. Little of this work, however, approaches AI's explainability problem from the physicians' point of view, *i.e.*, considering physicians' domain-specific needs and day-to-day practices.

Our goal is to enable physicians to understand existing data-driven, AI-enabled medical imaging analysis. We choose imaging as it is the primary data source in medicine [35]. In this paper, we focus on one of the most common modalities—chest X-ray (CXR) images. To achieve this goal, as shown in Figure 1, we started with a survey, which informed a user-centered design to formulate CheXplain—a system for physicians to interactively explore and understand CXR analysis generated by a state-of-the-art AI [32].

Iteration #1: Survey We conducted a paired-survey on both referring physicians (N=39) and radiologists (N=38) to learn about how radiologists currently explain their analysis to referring physicians and how referring physicians expect explanations from both human and (hypothetical) AI radiologists. The findings reveal whether, when, and what kinds of explanations are needed between referring physicians and radiologists. By juxtaposing referring physicians' response to these questions with respect to human vs. AI radiologists, we elicit system requirements that encompass the current practices and a future of AI-enabled diagnosis.

Iteration #2: Low-fi prototype. A co-design with three physicians manifested the survey-generated system requirements into eight specific features: augmenting input CXR images with specific inquiries, mediating system complexity by the level of urgency, presenting explanations hierarchically, calibrating regional findings with contrastive examples, communicating results probabilistically, contextualizing impressions with additional information, and comparing with images in the past or from other patients.

Iteration #3: High-fi prototype. We integrate the eight features into CheXplain—a functional system front-end that allows physician users to explore *real* results generated by an AI [32] while all the other explanatory information either comes from real clinical reports or is manually generated by a radiologist collaborator. An evaluation with six medical professionals provides summative insights on each feature. Participants provided more detailed and accurate explanations of the underlying AI after interacting with CheXplain. We summarize the design and implementation recommendations for future development of explainable medical AI.

Contributions
Our contributions are as follows:

- An empirically-informed, user-centered iterative design of CheXplain—a proof-of-concept system prototype that integrates eight key features for enabling physicians to explore and understand AI-generated medical imaging analysis;

- A summary of design and implementation recommendations for future development of explainable medical AI, encompassing four recurring themes: motivation, constraints, explanation, and justification. For example:

  - Insight: often what referring physicians look for is not explanation, but *justification*—information extrinsic to a CXR image that can support an AI's analysis result.
  - Recommendation: more explanation for medical data analysis (*e.g.*, radiologists), and more justification for medical data consumers (*e.g.*, referring physicians).

BACKGROUND: AI AND RADIOLOGY
The advent of data-driven AI has given rise to a plethora of research and development using AI to perform radiographic analysis [45, 55, 60]. The recent work on CheXpert [32] developed a deep neural network with a dataset that contains 224,316 chest X-ray images from 65,240 patients with 14 different observations. Such an 'AI radiologist' takes single CXR image as the input and outputs the labels that consist of a list of observations. The model in CheXpert has achieved a radiologist-level performance—an accuracy of 83% based on ground truths from CXR reports—and importantly, much faster diagnosis than human radiologists.

Nomenclatures
- Referring physician: a non-radiologist physician who sends imaging data (*e.g.*, CXR images) of a patient to a specialist for more information or treatment.

- Observation: in radiology reports, an observation is a description or finding from the radiograph.

- Impression: in radiology reports, impression comes as a list of summary statements or differential diagnoses concluded from the observations.

RELATED WORK
We review explainable AI (XAI) and its recent development in the HCI community. We then introduce literature in Clinical Decision Support Systems (CDSS), and since our focus is on imaging, we further discuss prior work on the communication between radiologists and referring physicians.

Explainable Artificial Intelligence (XAI) Systems
Researchers have focused on the explanations of expert intelligent systems since the 1970's [63]. The topic of explainability was rejuvaneted as recent developments in data-driven AI (*e.g.*, deep learning) exacerbates the obscurity of the hidden intelligent process. Ras *et al.* summarized recent XAI research into four topics: users, laws/regulations, explanations and algorithms [49]. A plethora of work on interpretable machine learning has sought to explain the inner principles of

the machine learning models with mathematical and algorithmic solutions [7]. However, there is no clear agreement on the definition of interpretability and controversy still exists. Lipton [40] proposed the motivations and methods to achieve interpretability. Doshi-Velez and Kim [19] discussed how to measure interpretability. Miller [45] defined interpretability as the degree to which a human can understand the cause of a decision and provided an overview of research about how people define, generate, validate, on present explanations.

To explain data-driven AI, researchers have focused on three categories of XAI solutions: rule-extraction methods, attribution methods, and intrinsic methods [49].

- Rule-extraction methods generate rules based on input and output information to simulate the âĂŸblack box' decision-making process. For example, Frosst and Hinton distilled a more understandable soft decision tree from a neural network [26]. Murdoch and Szlam constructed a simple, rule-based classifier that approximates the output of LSTM [33]. Ribeiro *et al.* developed LIME, which learns the model locally around the prediction and explains the prediction in an interpretable and faithful manner [51].

- Attribution methods measure the importance of a component to the output by changing the input or internal components. DeepLIFT decomposed the output prediction of a specific input by backpropagating the contribution of all the neurons to the input features [57]. Zintgraf *et al.* proposed a method of visualizing the response of a DNN to a specific input to analyze the prediction difference [71].

- Intrinsic methods enhance the interpretability of a model's internal constructs without affecting its performance. Santoro *et al.* transformed a DNN model into one that is able to answer rationale questions by adding a reasoning module that learns a relational function [52]. Others used a framework that consists of an attentive encoder-decoder network for video caption generation and an interpretive loss to visualize the learned features [18, 65].

Besides the works in ML and AI communities, in the HCI community, people attempt to understand AI from the users' perspective, which we review below.

Explainable Artificial Intelligence (XAI) Research in HCI
Intelligent sensing systems need to provide users with not only results but also accounts for their behaviors [5]. XAI in HCI contains topics that intersect context awareness, cognitive psychology, and software learnability [56]. For explainable context-awareness, the key is using simplistic representations of the context to inform users what is obtained and which action will be done by the systems [20]. Dey and Newberger designed a tailored interface that provides visual and textual explanations for context-aware rules [14]. Cognitive psychology focuses more on explanation theory. Lombrozo studied cognitive explanations and found that it is strongly connected with causality reasoning [41]. Software learnability is an important part of usability [28] and Abdul summarized topics of learnability such as hints, guidance as well as visualizations that are related to design an XAI system [1]. Another study proposed

a framework emphasizing empirical application-specific investigations of XAI by exploring the theoretical underpinnings of human decision-making [64]. Others focused on information visualization and ML algorithms visualization to help non-expert users understand AI [9].

However, the research in the HCI and AI communities often seem to be disconnected [1]. There is a lack of research that crosses and combines both fields to interdisciplinarily approach the XAI problem.

Clinical Decision Support Systems (CDSS)
Clinical Decision Support Systems (CDSS) provide knowledge to enhance medical decision-making for physicians. CDSS is widely adopted in fields such as screening and prevention, medication decision-making, therapeutic planning, and diagnostics [9]. With the advent of available medical data and data processing techniques, CDSS have become increasingly powerful in extracting useful information from a large amount of patient data to assist medical decision making [22, 46]. AI-based Medical Diagnosis Support Systems (AIMDSS) have become an important part of CDSS. Currently, research in AIMDSS primarily focuses on pathology and radiology, with the ability to detect nuanced patterns that are otherwise hardly noticeable by human doctors. For example, Jha and Topol mentioned AIMDSS is able to identify radiographs for radiologists and recognize patterns for pathologists to work as a information specialist during the diagnostic process [34].

However, the actual usage of these AI-based Clinical Decision Support Systems is limited because of a lack of understanding. Fan *et al.* found that initial and performance expectancy both have significant effects on doctors' behavioral intention of using AIMDSS [21]. Medical professionals may not use a system if the system cannot provide relevant information or capture the mental model of doctors [36, 38, 66]. Thus it is crucial that CDSS are able to explain themselves. The explanation capabilities of AI systems using knowledge bases were first added to medical decision making and computer-aided diagnosis in 1983 [59]. MYCIN was one of the earliest systems that incorporated domain knowledge and rule representations to achieve explainability [63]. DXplain explained and justified the interpretations by proposing knowledge-based diagnostic hypotheses including signs and symptoms to users in medical decision-making [4]. A diagnostic reasoning theory is proposed to find the components of the system that explain the discrepancy between the expected result and observed behaviors [50]. An argumentation-based interaction [25] that is flexible and easily understood by human users can help doctors make decisions based on this question. Better explanations can let users better understand the reasoning chain and enhance the system's confidence [9]. Park and Han proposed explanations as one of the methods of evaluating the clinical performance and effect of CDSS [48].

New research also emerged to address explainability challenges in medical AI systems that started to incorporate data-driven models. Caruana *et al.* developed an interpretable system using the generalized additive models with pairwise interactions GA^2M model predicting pneumonia risk and hospital 30-day readmission [11]. Krause *et al.* designed a visual

analytics system to help support interactive dependence diagnostics by feature representation and visualization [39].

In sum, research in CDSS attempted to bridge the gap of actual use by making it understandable to domain experts in medical decision-making. As this paper is conerned with XAI in radiology, below we review past research on the exchange of information between radiologists and referring physicians to understand their decision-making process.

Communication between Radiologists and Physicians
One study mentioned that the ambiguous requests from referring physicians may prevent radiologists from assisting physicians efficiently and a more precise medical history should be helpful for radiologists to better understand the cases [37]. Fischer *et al.* found that direct communication, informative request forms and questioning the patients will also be useful to lead to better communication between referring physicians and radiologists [23]. Another study proposed specific suggestions on how to improve reporting performance from referring physicians' points of view. The study indicated that reporting should receive more attention in training and practice than it currently does [29]. The American College of Radiology (ACR) also defined practice guidelines for communication of diagnostic imaging findings for radiologists aiming for better communication [2]. In addition, the quality of requests, transparency with regard to waiting time, a portal for easier scheduling, and making radiation dose information visible were also mentioned in a survey [53]. One study pointed out that the vagueness of the radiology report, *e.g.*, inappropriate use of words in radiology reports, would lead to information asymmetry between referring physicians and radiologists, which might prevent timely treatments [62]. Clinger mentioned referring physicians' satisfactory is mainly affected by diagnostic accuracy, clarity of language, and a detailed discussion of the findings [12]. Berlin pointed out that radiologists' duty extends beyond interpreting and sending out a written report to referring physicians. The communication of diagnosis is altogether as important as the diagnosis itself [6].

Despite all this research, little has been studied about how explanation—a specific form of communication—is currently used between referring physicians and radiologists, not to mention what difference there will be in an AI-enabled scenario. To bridge this gap and to inform the design of our system, below we started with a survey of explanations between referring physicians and radiologists.

ITERATION #1: LEARNING ABOUT EXPLANATIONS BETWEEN PHYSICIANS & RADIOLOGISTS
Given the limited prior art on explanation between physicians and radiologists, we conducted a survey to set the scene for the subsequent design activities. Specifically, we investigated whether, when, and what kinds of explanations are needed between referring physicians and radiologists when performing a diagnosis based on CXR images.

Survey Design
The authors all have computer science backgrounds. To understand communications between radiologists and referring physicians in real workflow, and to design survey questions, we conducted several interviews with two radiologists.

We employed a paired-survey design where both referring physicians and radiologists would answer the same set of questions, respectively, as explanation involves both an explainer and an explainee. For example, physicians would answer "what kinds of information do radiologists use to explain their examination results to you? ", whereas the same question for the radiologists would be "what kinds of information do you use to explain your examination results to the referring physicians?". This design allows us to compare and synthesize both parties' response *vis-Ãă-vis* in order to obtain a more complete picture of explanations between referring physicians and radiologists.

Further, we asked another set of questions only to referring physicians, specifically about when and what kinds of explanations they expect if the results were to come from an AI rather than a human radiologist. Given that very few AI-enabled radiograph analysis tools exist in clinical use, this set of questions is by nature speculative. Based on the state-of-the-art development [32], we described an AI radiologist to participants as a computer program that can produce—in real time—observations commonly found in human radiologists' report. This design allows us to identify if there is any difference in referring physicians' expectations of explanation when it comes to AI.

Participants
We distributed our surveys via social media groups of medical professionals across the US and personal connections to local medical centers. We recruited 77 participants in total (39 referring physicians and 38 radiologists). The study was conducted anonymously on Survey Monkey Platform and each participant received $20 as payment.

Findings
Whether Explanation Is Needed
In general, a lack of information motivates but the time cost constrains the seeking of explanations. Referring physicians often find themsevles in need for more information, *e.g.*, "*more comparisons*", "*more details instead of vague descriptions of abnormalities*". Calling radiologists or colleagues for explanation is considered "*one effective way to obtain information*". In the meantime, as is similar to the process of reporting [8], seeking explanation is also time-consuming: participants reported that waiting time for initiate a communication ranges from 15 minutes to one day, where only 56% of the participants felt satisfied with such efficiency.

With AI, a lack of trust becomes the motivating factor. Referring physicians remain skeptical, as only $1/3$ of them expressed trust of an AI radiologist, compared to $38/39$ reported likely or very likely to trust a human radiologist. Participants indicated that an AI radiologist would need to explain itself with "*probabilities*" and "*professional knowledge*".

When Explanation Is Needed
Referring physicians seek explanations when their questions are unanswered or when the results are atypical or critical.

Results show that asking for explanations of a specific question is one of the most frequent scenarios when referring physicians contact radiologists. As explainers, radiologists reported that they would like to "*know what the referring physician is looking for*" and "*have a more focused question rather than a vague description*" in order to generate more efficient and more specific explanations. Meanwhile, referring physicians looked for explanations for critical or atypical findings, *e.g.*, "*descriptions indicating high or low incidence of being significant*".

With AI, physicians would seek explanations when their hypotheses are in discordance with AI's results. Referring physicians reported more likely to need explanations when they believe they are right and AI's results are wrong or when they are not sure who is right.

What Kinds of Explanation Are Needed
Comparison with patients' prior images or other image modalities (e.g., CT scan) is most commonly used for explanations, as reported by both referring physicians and radiologists. Further, 18 participants considered more differential diagnoses as another way of explaining CXR-based results.

With AI, referring physicians further expect annotations on CXR and regional comparisons across patients. Referring physicians requested more than what current AI's textual description of the observations, *e.g.*, "*... area of concern instead of ambiguous descriptions*".

The needs for justifications emerged under this question. Different from explanation that explains with intrinsic process about how the results are generated, justifications apply extrinsic sources of information such as prior images or other image modalities to justify AI's results.

ITERATION #2: LOW-FI PROTOTYPING TO FORMULATE KEY SYSTEM DESIGNS
Building off of the survey's findings that point to system requirements, we conducted a user-centered design to formulate CheXplain in two major iterations. In this section we describe the initial low-fi prototype co-designed with three physicians.

Participants
Recruited from a local medical center, the three physicians came from different specialties but all had experience reading radiographs. As shown in Table 1, the participants had various degree of radiological knowledge, thus allowing us to calibrate system designs that serve for both novices and experts in radiology.

Apparatus & Procedure
To help participants understand the current status of 'AI radiologists', we first presented a simple 'system' based on the type of CheXpert's output without any explanation (Figure 2). Next, we used a combination of paper/pencil and Google Slides to create a series of low-fi prototype mock-ups. We used these artifacts as probes to elicit participants' reactions, critical feedback, and more discussions. Further, we asked participants to brainstorm more interactive features that could help them explore and understand an AI radiologist.

Participant	Specialty	Radiological Knowledge
P1	Pulmonologist	Competent
P2	Radiologist	Expert
P3	Allergist	Novice

Table 1. Iteration #2 physicians' knowledge in radiology. Novice: do not know much about radiology and do not report CXR images in workflow. Competent: know much about radiology but do not report CXR images in workflow. Expert: proficient in radiology and report CXR images in workflow.
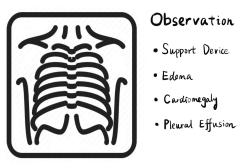

Figure 2. An example of AI-radiologist without any explanation to help participants understand the current status of AI radiologists

Results: Key System Designs
Through iterative development with three physicians, we consolidated the following key features of CheXplain. According to the findings in iteration #1, feature #1, #2, #8 are explanations and the others are justifications.

#1 Augmenting CXR images with specific inquiries
Prior AI-related work often considered a medical image as the only input and attempted to exhaust all possible diagnoses. However, as indicated in the survey and another paper [8], physicians often have a specific question for a CXR image, sometimes with important contextual information. P2 mentioned that "*For chest pain, they're usually looking for pneumothorax but for swallowing an object, they are looking for that object*". P1 and P3 suggested that allowing for a range of questions/auxilliary information as additional input.

Design As shown in Figure 3a, we design an initial step for referring physicians to select or type in keywords that characterize their questions specific to the patient's case. Accordingly, an AI should use such keywords to 'filter' the examination process, *e.g.*, only paying attention to relevant regions of the CXR image.

#2 Presenting explanations hierarchically
Previous work also mentioned that radiographical reports need to be itemized and better structured [8]. According to our discussion with physicians and existing radiologists' guidelines, there are three levels of radiographical information: *(i)* low-level examination—which body part to look at in a CXR image, *e.g.*, lungs, trachea; *(ii)* mid-level observation—what characteristics or phenomena are observed, *e.g.*, cardiomegaly, pleural effusion; *(iii)* high-level impression—differential diagnosis based on observations, *e.g.*, pleural effusion suggest pericardial disease or congestive heart failure.
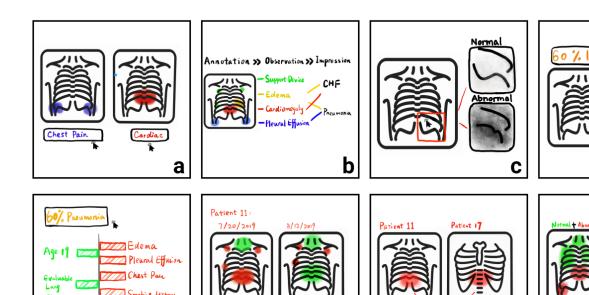
Figure 3. Low-fi prototypes of eight key system designs formulated in iteration #2. (*a*) Augmenting CXR images with specific inquiries: directed explanation based on a user's target question; (*b*) Presenting explanations hierarchically: connecting low-level annotation, mid-level observation, and high-level impression; (*c*) Contextualizing observations with contrastive examples: contrasting a user-selected region with both normal and abnormal samples; (*d*) Interpreting observations probabilistically: calibrating the understanding of AI's probability by referring to same observations in other CXRs with different probabilities; (*e*) Contextualizing impressions with prevalence & traces: presenting prevalence and known factors that affect a user-selected impression; (*f*) Comparison across time: showing the same patient's prior images; (*g*) Comparison across patients: comparing with CXRs from different patients having the same observation; (*h*) Mediating system complexity by urgency level: showing different amount of information in *urgent* vs. *non-urgent* mode.

Given such a hierarchy of information, we observed that more radiologically knowledgeable physicians tended to reason about AI's explanations in a bottom-up way while less knowledgeable physicians often took a top-down path when trying to make sense of AI.

Design As shown in Figure 3b, the main interface of CheXplain embraces the three-level hierarchy: annotations on the CXR image indicate low-level examinations (where the AI is looking at); observations and impressions are presented next to one another. As a user selects a specific examination/impression/observation, the system 'connects the dots': for example, by clicking *Pneumonia*, the user can see that AI makes that impression based on *Edema* observations, which are based on the examinations from CXR.

#3 Contextualizing observations with contrastive examples
Physicians mentioned that comparing abnormal and normal CXR images is a common practice in teaching radiologists. Relatedly, XAI research has proposed contrastive [15] and counterfactual [31] explanations that allow a person to understand something in light of plausible alternatives.

Design As shown in Figure 3c, as a user selects a specific region of examination on the CXR image, the system shows a pair of contrastive examples also generated by the same AI examining the same region. By looking at images that contrast the AI's observation, a physician can explore 'what-if'

questions, *e.g.*, "what if there is no cardiomegaly?" By looking at images of the same observation, a physician can see if there is any common pattern, and whether the current image shares that pattern (if not, it might be a misdiagnosed outlier).

#4 Interpreting observations probabilistically
Many AI models can output a probability for each possible conclusion, which can be leveraged to explain the result. For example, a seemingly unlikely observation/impression is explainable if there is a low probability. However, physicians expressed confusion about interpreting a numeric probability, *e.g.*, "*Why AI is only 90% sure, why is it not 100% sure? I would like to see the gap between 90% and 100%*" (P2). According to participants, we found that it was not the numeric values of probabilities that matter but how these values are mapped in medical workflows.

Design As shown in Figure 3d, as a physician user clicks an observation label, the system shows an array of CXR images of the same observation, sorted by their probability and calibrated by medical professionals' understanding of an AI observation's probability: unlikely (0%), likely (50%) and very likely (100%). For example, a 80% edema turns out to be quite similar to a near 100% case.

#5 Contextualizing impressions with prevalence & traces
Participants brainstormed various ways to support the understanding of impressions: P1 suggested "*using epidemiologic*

6

standpoints" as a baseline to explain the probability of an impression; P2 addiotionally suggested "*including the factors that AI took into consideration when generating the probability of impressions (e.g., the patient's age, risk factors, smoking history)*".

Design As shown in Figure 3e, the system provides two additional information to contextualize an impression: *(i)* prevalence, *i.e.*, prior probability of the same impression in a given population; and *(ii)* traces that go back to observations that contribute to higher or lower probability of such impression.

#6 Comparison across time

All three participants reacted positively to our initial idea of comparing with the same patient's prior images. P1 mentioned that an AI's observation is more explainable if the patient had similar diagnoses before. Similarly, P2 would like to see if the observations were already there in the prior images, and whether they were getting better or worse. P3 mentioned that previously non-existing observations should be prioritized when seeking explanations.

Design As shown in Figure 3f, a physician user can switch to a side-by-side view between the current and prior cases, view annotations and filter to see only what is different.

#7 Comparison across patients

To respond to the initial inquiry, participants suggested showing other patients whose cases had similar inquries.

Design As shown in Figure 3g, in the same view as comparing with prior images, a physician user can see a lists of other patients' cases responding to cardiomegaly inquiry.

#8 Mediating system complexity by urgency level

Our survey shows that time cost contrains referring physicians from seeking explanations. While an explainable AI radiologist can mitigate such problem, sometimes physicians might not even have time to interact with our system. P3 mentioned, "*I will just pick the important abnormalities if I am in a hurry and look back to see other information later*". P1 suggested two different modes of the system for urgent and non-urgent cases.

Design As shown in Figure 3h, we design an 'urgent' mode that can be toggled by the user, where the system only shows annotations to explain the most significant AI-generated observations, *i.e.*, observations that have both high confidence *and* lead to impressions of critical nature.

ITERATION #3: EVALUATING A HIGH-FI PROTOTYPE

We integrated the eight key features into a high-fidelity prototype of CheXplain, which allowed us to further iterate the design by having physicians interact with the system to explore and understand the AI-generated radiograph analysis.

Physicians

We invited 39 referring physicians who participated in the survey study and other medical professionals via connections to a local medical center. We finally recruited six participants. As shown in Table 2, the participants had various degree of radiological knowledge.

Participant	Specialty	Radiological Knowledge
P1	Pediatrics	Competent
P2	Internist	Expert
P3	Pediatrics	Competent
P4	Medical student	Novice
P5	Pediatrics	Competent
P6	Medical student	Novice

Table 2. Iteration #3 Participants' knowledge in radiology. Novice: do not know much about radiology and do not report CXR images in workflow. Competent: know much about radiology but do not report CXR images in workflow. Expert: proficient in radiology and report CXR images in workflow.

Apparatus

Figure 4 shows the study testbed system. We implemented the front-end of CheXplain as a web-based platform. All patient cases including input, regional annotations, prior and across-patient images were from the CheXpert dataset. Similarly, CXR observations were generated by a convolutional neural network used in CheXpert paper. The remaining requisite information was manually generated based on external medical guidelines [43] and was verified by a radiologist collaborator. Using manually-generated data to test a fully-implemented front-end allowed us to agily test the system design without committing to a computationally costly implementation of the back-end.

Physicians interacted with CheXplain using their own browsers. All the studies were conducted remotely via Zoom through which we observed physicians' interactions in real-time via screen sharing. We recorded the sessions with physicians' permission and transcribed the audio for further data analysis.

Tasks & Procedure

We started with a brief introduction of AI's application in medical image analysis. Next, physicians were presented with a patient's case randomly selected from the database, which came with textual labels indicating AI-generated observations. The main task was to interpret AI's results, *i.e.*, describing their understanding of why AI arrived at certain results of the case. We asked the physicians to respond as succintly and accurately as possible.

We asked each physician to describe their interpretation of AI after presenting CheXpert's results (Figure 4a). Then, we guided the physician to interact with CheXplain (Figure 4b), using it as a tool to explore and understand AI-generated results of the same patient's case. Afterwards, we asked the physician to describe their interpretation of AI again.

Physicians were also encouraged to think aloud during their interaction with CheXplain. We asked physicians *in situ* questions as they interacted with each of the eight key features— whether and how each feature helped them understand AI's results. We ended by asking each physician to summarize how CheXplain overall changed their understanding of the underly-
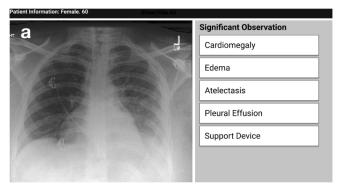
Figure 4. High-fi prototyping. (*a*) CheXpert: AI radiologist without explanations. (*b*) CheXplain: The high-fidelity prototype with 8 key features.

ing AI, and whether and how such systems can be integrated into their existing workflow.

Analysis & Findings
We analyzed each physician's response at a rolling basis. The first experimenter transcribed and coded physicians' responses and iteratively developed a code book. A second experimenter reviewed the codes and resolved disagreements via discussion with the first experimenter.

How physicians interpreted AI before & after CheXplain
We compare physicians' interpretation of AI before and after interacting with CheXplain. Results show that pre-CheXplain responses (*i.e.*, only relying on CheXpert's results) contain very little information—all but one physicians acknowledged that they were not able to provide an interpretation. The only physician who did respond described that "*AI gathered a lot of images with readings similar to human radiologists*", which was based on his previous knowledge of AI. In contrast, responses contained more substance after interacting with CheXplain. All physicians mentioned that AI arrived at the observations based on comparison against normal/abnormal images and prior images, which captured the data-driven, learning-based nature of AI. In addition, three physicians also mentioned that AI pointed out where it was looking at (P3, P5, P6)—"*It would give you a better idea because it tells you what areas to look at if you are not sure about the analysis and this is exactly what I would be looking for*" (P4).

Figure 5. Percentages of physicians' ratings for grouped features. The vertical axis presents groups in a top-down descending order by mean values of ratings.

Overall feedback: how CheXplain enables understanding AI
Participants were asked whether and how each feature helped them understand AI's result as well as rating the helpfulness from 1 (not helpful at all) to 5 (extremely helpful). We group the aforementioned eight design features into three different categories: image-based comparisons, interconnection among different information, and information filters. As shown in Figure 5, image-based comparisons were regarded as the most helpful features. Below we provide detailed qualitative discussion about each of the three categories (numbers in parentheses refer to individual features discussed in Iteration #2).

Image-based comparisons (#3, #4, #6, #7)
Image-based comparisons help physicians to understand AI by drawing on specific CXR examples as evidence, which helped physicians overcome the difficulty of understanding AI's results in isolation. Though image comparisons are common in image studies, CheXplain enables physicians to screening out whether AI makes mistakes in each step by comparisons. For example, *Interpreting observations probabilistically* by calibrating physicians' understanding of probability is helpful because offering images with different probabilities "*defines different likelihoods so I know which is statistically more likely*" (P1) while preventing users from "*missing something that's really obvious in AI's definitive category*" (P2). *Contextualizing observations with contrastive examples* was helpful by "*constrasting regions, I can tell what a normal region looks like, determine how similar the images are and understand AI's conclusion*" (P5). *Comparison across time* pointed out the difference of AI's observations from prior images, which might seem unlikely if viewed in isolation (P3, P6) and *Comparison across patients* surfaced common features for the same observation across multiple patients (P1, P2, P6).

However, there were confusions when physicians tried to understand AI-based on multiple sources of images. P1 and P2 thought AI also analyzed the patient's prior images when processing the present case, which suggests that design should clearly show what is and is not the input data to the AI. P3 mentioned that a common feature of the same observation might not always exist because sometimes it appears very differently across patients. To address this, future design should group cross-patient cases based on their variance so that physicians can look for common features only in the same group.

8

Interconnection among different information (#2, #5)
Although establishing interconnection of information is not novel in medical systems [27], we initially employ it in medical explainable AI. Participants found it useful for CheXplain to present AI's findings in an interconnected way, which serves as a road map for them to navigate CXR results at different levels of detail, although not all the interconnected information was found equally helpful for their understanding of AI.

For example, *presenting explanations hierarchically* provided an organized way for physicians to interpret AI's output, from which low-level annotations were found to be more helpful— "*I can know where AI looked at to figure out the result and it tells you exactly what's there, where, and why*" (P6). Impressions, on the other hand, were more controversial and often considered less helpful: P4 believed generating impressions is more complicated than what the system shows, which simply traces an impression to the lower-level observations and examinations. *Contextualizing impressions with prevalence & traces* was considered beyond the scope of the system. While physicians liked the concept of having such contextual information for each impression, most thought it was ambitious to explain both radiological and clinical factors, and were concerned that other important factors might be missing here. Future design can adopt an on-demand approach, *e.g.*, allowing physicians to ask specific questions about what they believe are relevant to contextualize a given impression.

Information filters (#1, #8)
Physicians found that both features in this category were not directly related to helping physicians understand AI; rather, physicians considered them as filters to help physicians target their understanding or reduce time cost of interacting with the system. Despite almost all AI research in medical domain claiming that they speed up the diagnosis, which are actually provided by computing power, we start from a physician-centered perspective and improve this problem by interaction.

Physicians mentioned that *Augmenting CXR images with specific inquiries* and *Mediating system complexity by urgency level* helped narrowing down the information when understanding AI and sorting significant observations fast—"*they help me to focus on observations that AI thinks are more likely to be related to my input questions*" (P5). P3 noted that AI presents findings related to his question input in the image.

Mediating system complexity by urgency level opened up physicians' various preferences in filtering information: P4 would like to "*see the important observations faster*" when running out of time; P6, on the other hand, "*still wants to see everything to get a better understanding*"; P3 preferred to "*pick what I think is urgent*" (P3). Further design could address these diverse requirements by making this feature customizable based on physicians' personal preferences.

DESIGN RECOMMENDATIONS FOR MEDICAL AI SYSTEM
Based on our survey and a user-centered process of prototyping CheXplain, we summarize design recommendations for developing medical AI systems that can be understandable to physicians. Specifically, there are four recurring themes across our three iterations of design activities: motivation vs. constraint and explanation vs. justification.

Motivation vs. Constraint
Motivation is what drives a physicians to try to understand an AI's diagnosis results, and constraint is what prevents them.

Enable a physician to achieve an understanding of AI targeted at a specific clinical problem, rather than to attain a general, open-ended understanding of AI. Our survey indicates that a lack of information motivates physicians to seek explanations; however, when co-designing CheXplain, physicians tended to focus on understanding aspects of AI's results only if they are relevant to their own hypothesis or differential diagnosis. Besides, in radiographic reporting, referring physicians also need to state clearly what clinical question they want to have answers to [8]. To address this, we recommend starting with a question input to narrow the scope of AI's analysis. Such a question-answering (QA) ability also presents implementation challenge to the medical AI community given the recent development of visual QA [3, 47, 54]. Currently, one potential pitfall is attempting to explain too much, or too deep into the technical stack, which might actually constrain a physician's ability to see and understand the more clinically relevant aspects of AI.

Make a wide range of data sources available as a physcian tries to understand the AI, as differential diagnosis seldom relies on a single type of data [67]. We recommend going beyond how a medical AI currently handles one specific type of input modality only (*e.g.*, CheXpert only looks at CXR images). A mixed-modality approach is more familiar to physicians and more likely to motivate them to understand AI's results. In contrast, one potential pitfall is developing multiple XAI systems for multiple types of data, *e.g.*, System A explains mining eletronic health records while System B explains deep convolutional network' processing of MRI data, which is likely to constrain a physician's effort in understanding how a clinical case is handled by AI.

Let physicians customize how much time they want to spend trying to understand AI. CheXplain provides an urgency toggle for physicians with low time budget to understand AI; yet this design still cannot address different approaches physicians take to juggle time and information. We recommend that, instead of adjusting time, the system should let a physician freely control the amount of information they would like to process, from trying to make sense of only the critical results, to revealing more explanation hierarchically, and to referencing other regional, previous or cross-patient examples.

Explanation vs. Justification
Our research started with a focus on explanations of AI, which enables physicians to understand AI with *intrinsic* processes of producing certain output given certain input data, *e.g.*, processing pixels of a CXR image to arrive at certain observations. Later, as we co-designed CheXplain with physicians, a different set of ideas emerged, which tried to draw on *extrinsic* sources of information to justify AI's results. Such justifications include prevalence of a disease, contrastive examples

from other CXR images, and a patient history. Thus, we consider explanations and justifications as two categories of methods that help physicians understand AI.

Use more explanation for medical data analysts (*e.g.*, radiologists), and more justification for medical data consumers (*e.g.*, referring physicians). We find that radiologists tend to expect more explanations for details such as low-level annotations of CXR images, while referring physicians are generally less concerned about the instrinsic details but care more about extrinsic validity of AI's results. Retrospectively, we can see that five of CheXplain's features are justification (#3-7).

Enable explanation and justification at different levels of abstraction, similar to how CheXplain employs the examination-observation-impression hierarchy to scaffold both explanation and justification. Holistically, as a physician follows a bottom-up or top-down path to understand an AI's diagnosis of a patient, at any step along the way, they should be able to seek both explanation and justification. To achieve this, the XAI community need to consider explanation regulated by a user-specified level of abstraction; research on Content-Based Image Retrieval (CBIR) should enable search criteria at multiple levels of abstraction, *e.g.*, from a region of specific visual features to a global image that presents a specific impression.

Implementation challenges for medical imaging AI: explain how AI 'looks at' a region; justify how two images are clinically (dis)similar.

A plethora of prior work [13, 24, 54, 58] has achieved localizing where AI is 'looking at' on an input image, but not *how*—*e.g.*, how does AI consider this region as edema? Generating a natural language explanation—analogous to Hendricks' work on bird classification [30]—presents a new challenge for medical imaging AI.

When comparing images to justify AI's result, one challenge is to inform physicians how two images are radiographically (dis)similar, thus further enable them to gain insight from such justifications. Cai *et al.* designed a TCAV-like [10] tool for pathologists to find images based on self-defined concepts, which should be generalized to enable justifications beyond comparing visual features.

LIMITATION
Our current work has the following limitations:

- Survey questions regarding an AI radiologist were inevitably speculative due to the limited adoption of AI in current medical practices;

- The number of participants was limited due to the scarcity of medical professionals' availability;

- The available CXR dataset we used has limited information to further the exploration using CheXplain, *e.g.*, missing patients' medical history, no additional lab results.

- The final evaluation of system design is inevitable speculative instead of in a real workplace due to strict regulations.

The results may have slight differences in the actual diagnostic workflow.

CONCLUSION
In this paper, we identified the gap between current XAI research and domain experts' understanding of AI in medical decision-making. Our work contributes to a bridge of this gap by conducting a physician-centered design of CheXplain — a system that enables physicians to understand AI-enabled chest X-ray analysis by interaction. Ultimately, we propose a series of recommendations for future human-centered medical AI development around four recurring themes: motivation vs. constraint and explanation vs. justification.

Taken together, our work provides implications for how physicians can explore and understand data-driven, AI-enabled medical imaging analysis to assist physicians in medical decision-making process. The future work will explore the development of human-centered medical AI beyond medical imaging analysis.

REFERENCES
[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 582:1–582:18.

[2] Acr. 2010. ACR Practice Guideline for Communication of Diagnostic Imaging Findings. 1076, Revised 2008 (2010), 1–6.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[4] G O Barnett, J J Cimino, J A Hupp, and E P Hoffer. 1987. DXplain. An evolving diagnostic decision-support system. *JAMA* 258, 1 (July 1987), 67–74.

[5] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* 16, 2-4 (Dec. 2001), 193–212.

[6] Leonard Berlin. 2007. Communicating results of all radiologic examinations directly to patients: has the time come? *AJR Am. J. Roentgenol.* 189, 6 (Dec. 2007), 1275–1282.

[7] O Biran and C Cotton. 2017. Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI)* (2017).

[8] Jan ML Bosmans, Joost J Weyler, Arthur M De Schepper, and Paul M Parizel. 2011. The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys. *Radiology* 259, 1 (2011), 184–195.

[9] A Bussone, S Stumpf, and D O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169.

[10] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, and others. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 4.

[11] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1721–1730.

[12] Neal J Clinger, Tim B Hunter, and Bruce J Hillman. 1988. Radiology reporting: attitudes of referring physicians. *Radiology* 169, 3 (1988), 825–826.

[13] Piotr Dabkowski and Yarin Gal. 2017. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*. 6967–6976.

[14] Anind K Dey and Alan Newberger. 2009. Support for context-aware intelligibility and control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 859–868.

[15] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[16] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Jan. 2016), 56–62.

[17] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning*. jmlr.org, 647–655.

[18] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. 2017. Improving interpretability of deep neural networks with semantic information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4306–4314.

[19] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (Feb. 2017).

[20] Paul Dourish. 1995. Developing a reflective model of collaborative systems. *ACM Trans. Comput. Hum. Interact.* 2, 1 (March 1995), 40–63.

[21] 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). *Ann. Oper. Res.* (March 2018).

[22] M Fieschi. 2013. *Artificial Intelligence in Medicine: Expert Systems*. Springer.

[23] Harry W Fischer. 1983. Better communication between the referring physician and the radiologist. *Radiology* 146, 3 (1983), 845–845.

[24] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[25] J Fox, D Glasspool, D Grecu, S Modgil, M South, and V Patkar. 2007. Argumentation-Based Inference and Decision Making–A Medical Perspective. *IEEE Intell. Syst.* 22, 6 (Nov. 2007), 34–41.

[26] Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a Neural Network Into a Soft Decision Tree. (Nov. 2017).

[27] Dimitris Gritzalis and Costas Lambrinoudakis. 2004. A security architecture for interconnecting health information systems. *International Journal of Medical Informatics* 73, 3 (2004), 305–309.

[28] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: An Investigation of Contextual Video Assistance for Functionality Understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1515–1524.

[29] R Gunderman, W T Ambrosius, and M Cohen. 2000. Radiology reporting in an academic children's hospital: what referring physicians think. *Pediatr. Radiol.* 30, 5 (May 2000), 307–314.

[30] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*. Springer, 3–19.

[31] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809* (2018).

[32] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and Others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*. aaai.org.

[33] W James Murdoch and Arthur Szlam. 2017. Automatic Rule Extraction from Long Short Term Memory Networks. (Feb. 2017).

[34] Saurabh Jha and Eric J Topol. 2016. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* 316, 22 (Dec. 2016), 2353–2354.

[35] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2, 4 (Dec. 2017), 230–243.

[36] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Med Inform* 6, 2 (April 2018), e24.

[37] Charles Kilhenny. 1972. Improving communications between team physician and radiologist. (1972).

[38] Ajay Kohli and Saurabh Jha. 2018. Why CAD Failed in Mammography. (2018).

[39] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5686–5697.

[40] Zachary C Lipton. 2016. The Mythos of Model Interpretability. (June 2016).

[41] Tania Lombrozo. 2010. Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cogn. Psychol.* 61, 4 (Dec. 2010), 303–332.

[42] Tania Lombrozo and Susan Carey. 2006. Functional explanation and the function of explanation. *Cognition* 99, 2 (March 2006), 167–204.

[43] Rodrigues Mark and Qureshi Zeshan. 2017. The Unofficial Guide to Radiology: Chest, Abdominal and Orthopaedic X Rays, Plus CTs, MRIs and Other Important Modalities: Core Radiology Curriculum. (2017).

[44] D Douglas Miller and Eric W Brown. 2018. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am. J. Med.* 131, 2 (Feb. 2018), 129–133.

[45] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.

[46] Travis B Murdoch and Allan S Detsky. 2013. The inevitable application of big data to health care. *JAMA* 309, 13 (April 2013), 1351–1352.

[47] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757* (2016).

[48] Seong Ho Park and Kyunghwa Han. 2018. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 286, 3 (March 2018), 800–809.

[49] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven (Eds.). Springer International Publishing, Cham, 19–36.

[50] Raymond Reiter. 1987. A theory of diagnosis from first principles. *Artif. Intell.* 32, 1 (April 1987), 57–95.

[51] 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*. ACM, 1135–1144.

[52] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*, I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.). Curran Associates, Inc., 4967–4976.

[53] Sectra. 2013. How radiology can improve communication with referring physicians. June (2013).

[54] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[55] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua , and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2497–2506.

[56] Ben Shneiderman. 2016. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proc. Natl. Acad. Sci. U. S. A.* 113, 48 (Nov. 2016), 13538–13540.

[57] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 3145–3153.

[58] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[59] William R Swartout. 1983. *Xplain: A system for creating and explaining expert consulting programs*. Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

[60] R Takemiya, S Kido, Y Hirano, and S Mabu. 2019. Detection of pulmonary nodules on chest x-ray images using R-CNN. In *International Forum on Medical Imaging in Asia 2019*, Vol. 11050. International Society for Optics and Photonics, 110500W.

[61] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM Trans. Comput. -Hum. Interact.* 24, 3 (April 2017), 21:1–21:40.

[62] Carlos Valls. 2001. Pitfalls of the Vague Radiology Report. (2001).

[63] William van Melle, Edward H Shortliffe, and Bruce G Buchanan. 1984. EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project* (1984), 302–313.

[64] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. (2019).

[65] Tianfu Wu, Wei Sun, Xilai Li, Xi Song, and Bo Li. 2017. Towards Interpretable R-CNN by Unfolding Latent Structures. (Nov. 2017).

[66] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. *ACM Trans. Comput. Hum. Interact.* 2016 (May 2016), 4477–4488.

[67] Xie Yao, Ge Gao, and Xiang 'anthony' Chen. 2019. Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis. (Feb. 2019).

[68] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*. Springer International Publishing, 818–833.

[69] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2979–2989.

[70] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. 2019. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Med. Image Anal.* 56 (Aug. 2019), 80–95.

[71] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. (Feb. 2017).