Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust

Eleanor R. Burgess
Carelon Digital Platforms, Design & User Research, Elevance Health
eleanor.burgess@carelon.com

Ivana Jankovic
Carelon Digital Platforms, Point of Care AI
ivana.jankovic@carelon.com

Melissa Austin
Carelon Digital Platforms, Design & User Research, Elevance Health
melissa.austin@carelon.com

Nancy Cai
Carelon Digital Platforms, Design & User Research, Elevance Health
nancy.cai226@gmail.com

Adela Kapuścińska
Carelon Digital Platforms, Design & User Research, Elevance Health
kapuscinska.adela@gmail.com

Suzanne T. Currie
Chief Experience Office, Elevance Health
suzanne.currie@elevancehealth.com

J. Marc Overhage
Enterprise Analytics Core, Elevance Health
joseph.overhage@elevancehealth.com

Erika S Poole
Carelon Digital Platforms, Design & User Research, Elevance Health
erika.poole@carelon.com

Jofish Kaye
Carelon Digital Platforms, Design & User Research, Elevance Health
acm@jofish.com

ABSTRACT

Artificial intelligence (AI) supported clinical decision support (CDS) technologies can parse vast quantities of patient data into meaningful insights for healthcare providers. Much work is underway to determine the technical feasibility and the accuracy of AI-driven insights. Much less is known about what insights are considered useful and actionable by healthcare providers, their trust in the insights, and clinical workflow integration challenges. Our research team used a conceptual prototype based on AI-generated treatment insights for type 2 diabetes medications to elicit feedback from 41 U.S.-based clinicians, including primary care and internal medicine physicians, endocrinologists, nurse practitioners, physician assistants, and pharmacists. We contribute to the human-computer interaction (HCI) community by describing decision optimization and design objective tensions between population-level and personalized insights, and patterns of use and trust of AI systems. We also contribute a set of 6 design principles for AI-supported CDS.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Artificial intelligence, machine learning, type two diabetes, knowledge creation, provider workflows, sociotechnical complexity, medication prescribing, design principles, design objective

ACM Reference Format:

Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne T. Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3581251

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581251

1 INTRODUCTION

Over the past few decades, increasing amounts of health data have been aggregated by organizations (e.g., hospitals, health insurers) and individuals (e.g., patients reviewing continuous glucose monitors; athletes tracking smartwatch data). Data sources, including medical claims, electronic health records (EHRs), biometrics and lab tests, remote device monitoring, and genetic information, are known in the clinical domain as "observational data" or "real world evidence" (RWE)—stories of people's health choices, treatment(s), and health outcomes. RWE can support research at scale, however bringing many sources of rich data together is challenging. Artificial intelligence (AI) and machine learning (ML) tools can integrate these complex datasets to create insights. While much work is underway to determine the technical feasibility and accuracy of developing AI-driven insights, the specific insights that are considered useful and actionable by healthcare practitioners remain unclear.

Clinicians make treatment decisions based on their medical judgement, which may involve integrating professional experience and best-practice guidelines. Such guidelines are a consensus approach to a clinical problem based on clinical trials and expert opinion and may be specific to individual countries or even individual medical associations. Best-practice guidelines are updated with some regularity; for example, the American Diabetes Association updates its guidelines annually [54]. Given the way that best practice guidelines are currently constructed (or "designed"), introducing other forms of knowledge prompts interesting challenges. For instance: How do clinicians make sense of AI-driven insights in relation to other information they have on hand? And importantly, how might they trust these AI insights, given what they know about other ways of creating knowledge in the world

such as a clinical trial? As Beede et al. [4] state, medical organizations and practitioners will have a "difficult time dealing with AI if it does not integrate effortlessly into their present infrastructure, or even worse if it adds more complications."

Here, we study healthcare practitioners' reactions to an AI-based clinical decision support (CDS) prototype mockup designed to display insights guiding medication selection for Type 2 diabetes mellitus (T2DM). Our research goals were to deeply understand the workplace context and challenges of medication prescribing and to acquire participant feedback to iterate on the AI-based prototype concept. The purpose of the conceptual prototype was to provide prescribing clinicians AI insights regarding medication regimens for patients with poorly controlled T2DM, a condition with high prevalence in the U.S. population and high treatment costs, accounting for nearly a quarter of all healthcare spending in the United States [2]. Worldwide, an estimated 462 million individuals manage T2DM [36]. There are at least seven major drug classes for the treatment of T2DM, many of which can be taken at the same time [54]. The prototype aimed to analyze a wide set of possible treatments and present insights regarding potential effective treatments, especially combinations of medications, for which evidence is currently lacking in the clinical trial literature.

Using this T2DM treatment CDS could enable improved patient outcomes, identify patients who are under-managed, reduce time spent in trial-and-error processes to determine optimal medication combinations, and assist patients in switching to more effective treatments more rapidly. However, there is a gap between the ability to produce these insights and clinicians incorporating the insights as part of real-world patient care. Our study investigated the requirements of integrating an AI-powered CDS tool into the intense setting of the patient visit, where providers have minimal time, information-rich EHR tools but no time to dig deeper into the records, and may see over 900 technology alerts daily [37]. We iterated on the prototype design over the course of the study through 41 interviews with primary care and internal medicine physicians, endocrinologists, nurse practitioners, physician assistants, and pharmacists. The AI model underlying the studied prototype was based on a study involving N=141,625 patients [6], a significantly larger sample size than most clinical trials. The real-world validity of the study is powerful. What would be required for these types of insights to be accepted by the clinical community?

Critically, we found that AI systems for medication insights are judged against "gold standard" methods of clinical knowledge generation, especially randomized controlled clinical trials (RCTs). Within RCTs, trust is engendered in the precision of clinical trial processes and the standards by which they are carefully operated to avoid systematic bias. In this study, we evaluate how confidence or trust in the AI-supported insights may be influenced by the user's understanding and confidence in the methods used to generate the insight, a major focus of explainable AI work [58]. The prototype also elicited rich descriptions of clinicians' complex decision-making processes balancing efficacy, affordability, and patient preferences.

Responding to Andersen et al.'s [49] call to further understand "sociotechnical uncertainties" of AI systems in healthcare, we contribute to the HCI community by delineating implementation barriers for AI-supported CDS across the U.S. healthcare system, including decision optimization tensions between population-level and personalized insights, core responsibilities of the clinician and patterns of use and trust of AI systems. We also contribute six design principles for AI-supported CDS.

2 RELATED WORK

An interdisciplinary project like ours builds on a wide array of existing literature. Section 2.1 provides a short overview of Type 2 diabetes mellitus. In Section 2.2 we share recent research about Artificial Intelligence and Machine Learning Technologies in Human-Computer Interaction. In Section 2.3 we briefly overview HCI research regarding clinical workflows and healthcare practitioners' use of software tools. We conclude in Section 2.4 with relevant literature about Clinical Decision Support technologies from clinical and informatics domains.

2.1 Type 2 Diabetes Mellitus

Type 2 diabetes mellitus (T2DM) is a common chronic illness worldwide. More than 37 million people in the United States manage diabetes (about 1 in 10 Americans), and 90-95% of this population has T2DM [16]. Diabetes is characterized by high blood glucose (sugar) caused by a relative or absolute deficiency in the hormone insulin. The pancreas produces insulin to allow glucose to enter cells for energy [16]. If insulin is absent or a patient has insulin resistance, glucose remains in the blood at elevated levels. Chronic high blood sugar can eventually damage blood vessels and nerves, leading to complications such as heart attacks, kidney failure, limb amputations, and blindness [16, 45].

T2DM is diagnosed with a blood test, either by measuring the glucose level or checking hemoglobin A1c (A1c), which captures approximately a three-month window into a patient's glucose levels [15]. A patient's A1c is usually elevated in diabetes ($\geq$ 6.5%) and lowering A1c to a personalized target is a goal of diabetes management. Treatment for T2DM is based on lifestyle modifications but often requires additional oral and/or injectable medications.

The first-line therapy for T2DM is usually metformin, and additional medications are chosen based on a patient's other comorbidities, personal preferences (e.g., needle aversion), and disease severity. In more severe forms of the disease, it is often necessary to treat with insulin, an injectable medication that can be lethal if overdosed [19]. Because of the large number of second-line therapies and dearth of studies on the best combinations beyond one or two drugs, there is significant heterogeneity in treatment combinations for T2DM [28], suggesting the potential to improve personalized treatment regimens for many patients.

2.2 Artificial Intelligence and Machine Learning Technologies in HCI

HCI's longstanding interest in interactions with AI systems has continued to grow as machine-learning technologies have matured and been deployed outside pure research environments. Healthcare has remained a core domain for AI/ML-driven systems. We

focused on research from the last decade or so, corresponding to the growth in the use and implementation of deep-learning neural networks. For example, there have been several workshops on the topic of HCI in AI-based healthcare applications during this time: Realizing AI in Healthcare: Challenges Appearing in the Wild [49], Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare [51], and Patient-clinician communication: the roadmap for HCI [64], just as a few examples, each of which contained several papers exploring relevant issues. Similarly, the November-December 2018 issue of interactions included a Special Topic on Designing AI [30], drawing from the work presented at the 2017 and 2018 AAAI Spring Symposium on designing machine learning tools' experience [38]. A core area was the development of transparent, explainable, accountable, and intelligible systems [1], in part driven by DARPA's Explainable AI (XAI) program [59]. We also drew on prior work on AI-supported CDS systems deployed in the wild, such as [13].

In this paper, we focus on an area of study described by Vereschak et al. [61] as AI-assisted decision-making, where "humans make decisions based on their own expertise and on recommendations provided by an AI-based algorithm." Recent research has focused on the importance of onboarding clinicians through appropriate training to enable them to understand and collaborate with AI CDS [13, 14]. Specifically, Cai et al. [13] note that importance of sharing the known strengths and limitations of the AI-support system, the subjective point-of-view, and its overall design objective – "what it's designed to be optimized for." In this paper, we contribute to the concept of design objective, first by showing the challenge of a singular design objective and then by describing the list of multiple objectives clinicians consider when recommending T2DM medication.

Following Wang et al.'s [63] investigation of a diagnosis and medication recommendation AI-supported CDS deployed in rural China, they advocate for future tools to be designed to support a human-AI collaboration paradigm. They write, "the AI system should also be 'cooperative' – it can work together with human clinicians, fit into the local context, integrate with existing IT systems, and improve work productivity in the workflow." Additionally, they push for AI systems to follow the guidelines of shared decision-making framework taking patients' "social, cultural, and personal context into consideration in order to generate more personalized recommendations." These are essential yet challenging goals to meet and require rich context-specific research investigations to form a foundation for technologies which can facilitate these insights. In this paper, we explore how the US healthcare system context of diabetes care influences the use of an AI-supported CDS. We particularly highlight misalignments regarding personalization showing the deep complexity of this task deeply embedded in the "art of medicine".

2.3 Healthcare Provider Workflows and Electronic Health Record Use

A key technology mediating healthcare activities is the Electronic Health Record (EHR) [22], alternatively called electronic patient records or electronic medical records. In addition to digital patient records, EHRs also contain many additional modules, including e-prescribing, secure email, care management information, and care gap checklist dashboards. In their investigation of clinicians' use of an EHR system for diabetes care, Veinot et al. [60] describe several use categories: "priming, structuring, assessing, informing, and continuing." These categories show how the EHR is used both for preparatory work (e.g., reminding a clinician about the patient's recent medical history; preparing for what to ask the patient during the visit) and to orient the flow of questions, documentation, resource-sharing, and next-steps ordering during the encounter itself. The prototype investigated in our study was developed to integrate into an EHR system supporting Veinot et al.'s clinical consultation category of "informing".

HCI researchers have also investigated tools to support patient-clinician communication. Several studies [8, 9, 11, 39] have shown that patient-clinician communication during a clinical encounter is difficult, not least because of power and knowledge differences. As a first step to support collaborative reflection and learning, Mamykina et al. [41] created a tool called MAHI, a health monitoring application developed to assist newly-diagnosed individuals with diabetes in developing reflective thinking skills through social interaction with diabetes educators. Through prompted reflection and conversations over time with diabetes educators, patients were able to address self-management challenges and gain important self-efficacy in their diabetes management, moving toward an internal locus of control.

An ongoing challenge for clinical technologies is the severe time constraints on clinical decision-making. Many practitioners have minimal time with patients, sometimes as little as 10 minutes [42], and a primary care physician simply following guideline-recommended care would require more than 24 hours/day of work [53]. This time pressure challenges the recommended approach of shared decision-making (e.g., [25, 43]) where clinicians and patients discuss treatment decisions together.

Our study investigates the requirements of integrating an AI-powered clinical decision support tool into this intense setting, where providers have minimal time, have information-rich EHR tools but no time to dig deeper into the records, and may see over 900 technology alerts daily [37]. We add to the HCI literature by exploring the use of an AI-supported CDS, which introduces further open questions for clinicians regarding trust and explainability. We investigate how these questions intersect with clinician's goals within these time-constrained settings.

2.4 Clinical Decision Support (CDS) Systems in Healthcare Settings

Clinical Decision Support (CDS) systems have long been of interest to technology researchers. CDS is the process that "provides clinicians, staff, patients or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care" [47]. CDS systems support clinical decision-making by presenting insights based on patient-specific characteristics in a clinical scenario. CDS can support many aspects of the care process, including reminders for preventative care, diagnosis or treatment, or alerts related to health conditions. For example, a CDS tool can support the process that "provides clinicians, staff, patients or other individuals with

knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care" [47].

We think of CDS systems as incorporating patient-specific information about the clinical situation into an algorithm, such as a rules-based decision tree or a machine learning model and presenting the output as recommendations to the user. CDS can support many aspects of the care process, including prevention, diagnosis, treatment and monitoring of health conditions, addressing coding inaccuracies [5], and alerting the ordering clinician of a drug interaction [44]. CDS systems may be standalone software programs or integrated into EHRs; in either case, their ubiquity in clinical medicine raises interesting questions about how clinicians and other users trust such systems.

Creating effective CDS systems has proven challenging. Decades of experience have been crystallized into the five rights of CDS: "provide the right information, to the right person, in the right format, through the right channel, at the right point in the workflow" [48]. However, a poorly designed CDS has been shown to contribute to alarm fatigue, leading to overrides, workarounds, and clinician burnout [33]. Even more troubling, few studies of CDS have shown improvement in patient outcomes [26, 33]. Attaining the "five rights" remains aspirational for many CDS technologies [48].

Artificial intelligence-based clinical decision support has been around for decades [40, 57]. With improvements in computational power, it became possible to utilize extensive amounts of health data to train AI models for CDS on a reasonable timescale. However, compared to rules-based CDS (e.g., an alert that fires when a patient has a listed allergy to an ordered medication), the ethics and understandability of AI-based CDS is more complex. For example, the lack of explainability of some AI models can make it challenging for clinicians to understand the generalizability of CDS and confounding in the data can perpetuate hidden biases [40]. Indeed, AI models can sometimes function as "black boxes"; the "rules" they follow to make decisions are largely opaque [20]. This lack of transparency in how, precisely, a given AI model yields an output poses a trust issue for the clinician, who must decide if the CDS based on said model should be followed for their patient [20].

3 METHODS

Here we describe the prototype including its AI model and interface (see Section 3.1) and then our study participants (see Section 3.2), data collection (see Section 3.3), and data analysis and researcher position (see Section 3.4).

3.1 Type 2 Diabetes Mellitus Medication Insights Prototype

The machine learning model underlying the T2DM prototype was trained on a large health insurance claims dataset to optimize medication selection for hemoglobin A1c reduction [15]. Hemoglobin A1c (also referred to as A1c, HbA1c, or HgA1c), is a measure of diabetes severity. The purpose of the tool was to provide prescribing clinicians with AI insights regarding medication regimens most likely to reduce a patient's A1c based on similar patients. The prototype analyzed a rich set of treatment alternatives, including multiple drug combinations that, while individually clinically validated, have yet to be evaluated together in clinical trials, and presented insights regarding potential effectiveness.

3.1.1 Model Concept. The underlying model producing insights for the prototype used a dataset from medical claims data of nearly 150,000 patients with an A1c of $\geq$ 9%, and at least one follow-up HbA1c measure within 3-6 months [6]. An A1c $\geq$ 9% is considered "poor diabetes control" by the Center for Medicare and Medicaid Services [18]. The model cohorted patients by age, number of comorbidities, and whether the patients had previously been on insulin or not. Steps were taken to explicitly reduce the impact of data and selection bias in the model [7]. By comparing the A1c levels before and after these drug combinations were prescribed, the researchers could determine the relative efficacy of different drug insights compared to the baseline treatment of metformin alone, the popular first-line treatment for diabetes. The output of the model was a ranked list of medication treatment options optimized for A1c reduction in each cohort under consideration. The model was designed in accordance with Elevance Health's requirements on the responsible design of AI projects, including approval from the Office of Responsible AI.

3.1.2 Model Data. Medical data can reveal much about the real-world experiences of patients. Data such as the results of lab tests can help researchers understand the potential efficacy of a treatment for an individual patient. These data can present a temporal perspective: across a patient's illness trajectory, when and for how long did they receive these assessments and treatments? With substantially large datasets, potential effectiveness of interventions across populations ("cohorts") of patients can be determined. Given their large size and historical perspective (often delineating care across years of a patient's life), medical datasets can power insights at scale. It is important to note, however, that the structured data available in many medical datasets only captures a subset of the full patient experience. For example, it includes no unstructured data, such as notes written by a clinician about a patient encounter (known as case notes) or images, let alone information patients may convey explicitly or implicitly to the clinician that is not recorded in the case notes. As such, a model built on structured data is unable to account for information that might be captured in notes, images, or other unstructured data about patient behavior change or lifestyle modifications, such as changes in diet or exercise patterns.

3.1.3 Prototype. After the model was complete, a series of prototype interfaces were developed to showcase these insights to clinicians for feedback. These were non-interactive mockups shown to clinicians over Zoom. Iteration cycles were rapid and included sometimes daily changes to interface elements. Changes were made in the prototype in response to feedback from a small number of participants and stakeholders to rapidly home in on an optimal data display. The below images, showing design waypoints, illustrate diverse points in the evolution of the prototype. We discuss the prototypes below and share participant responses in Section 4.1 and 4.2. We began with Figure 1 (Prototype A), which surfaced the top alternate regimens to a simulated patient's current regimen, based on A1c change.

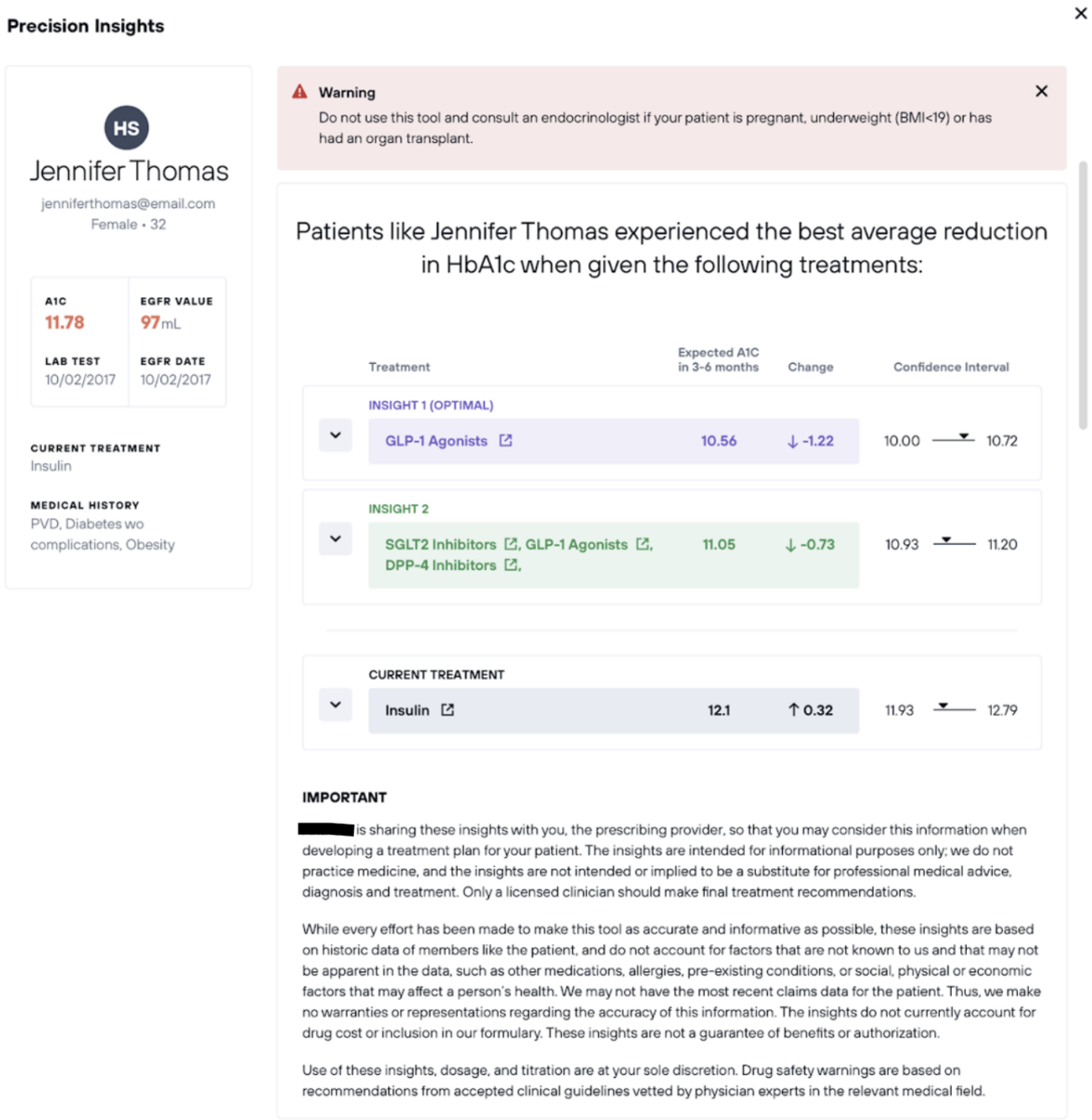

Figure 1: Prototype A. Patient data is simulated; outcomes are calculated by the model. Note that because the data is simulated, the order of treatment regimens depicted here may not align as well with clinical expectations as a live tool.

Prototype A has three main sections. The left margin shows core information about the (simulated) patient (e.g., name, contact information, sex, age) and two key lab tests: the A1c measurement and the eGFR value—a measure of kidney function—an important consideration for certain medication. It also includes the patient's current treatment and key points from their medical history.

The middle section of the prototype, with the title "Patients like Jennifer Thomas experienced the best average reduction in HbA1c

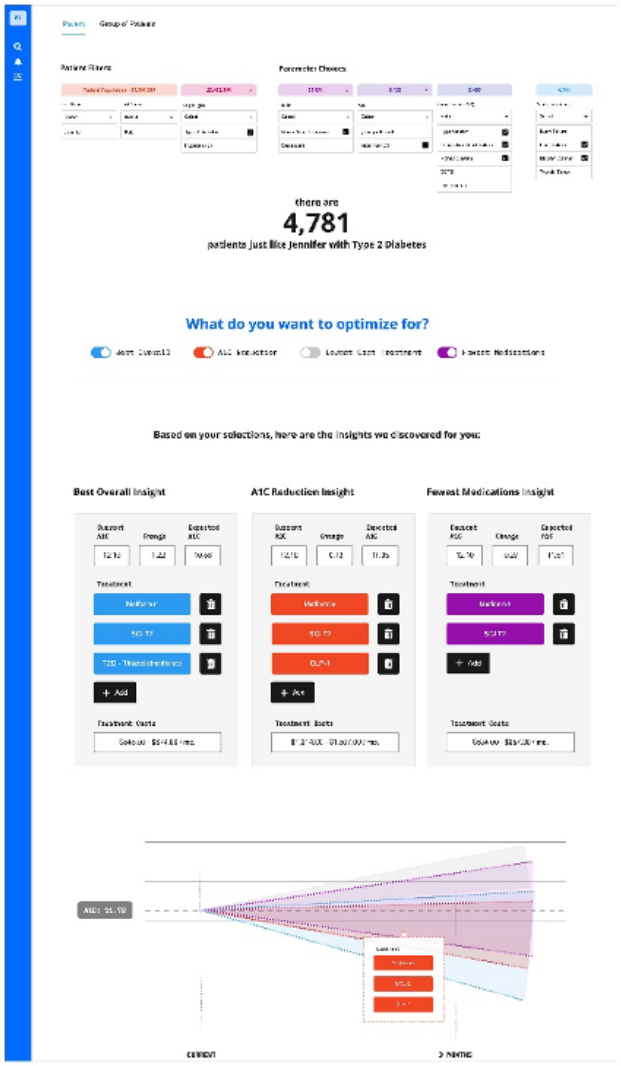

Figure 2: Prototype B. Patient data and cohort sizes are simulated; outcomes are calculated by the model. This sketch is from Iteration 8, which was focused on exploring different ways to show the uncertainty in the A1c change for a particular regimen.

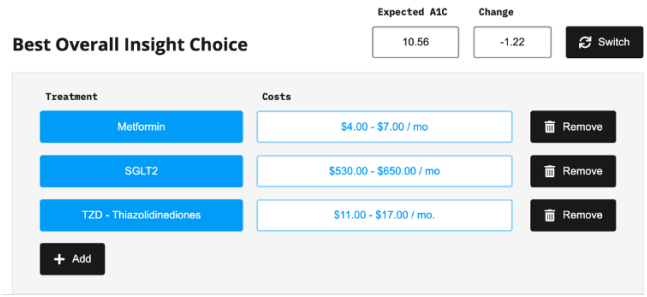

Figure 3: This popup screen, also from Prototype B, shows more detail, including estimated associated costs for individual components of the treatment. See Section 4.1 for a more detailed discussion of the role of costs in treatment decisions.

when given the following treatments:" proposes two alternate treatment regimens with best A1c results. In this case, a GLP-1 agonist, or a combination of SGLT2 inhibitors, GLP-1 agonists, and DPP-4 inhibitors. For each treatment the view lists relevant data: the expected A1c due to medication in 3-6 months, the change from the patient's current levels, and the confidence interval of that change. It also documents the patient's current treatment, showing the same data. Each of these sections can be unfolded by clicking to show more information about these treatment types, including warnings specific to those drug classes.

Finally, the top and bottom include warnings and legal disclaimers. Their presence both provided a place for the institution's legal team to weigh in on wording, as well as highlighting issues that clinicians—both on the model development team and participants in the study—would expect to see in the context of similar insights.

The prototype shown in Figure 2 (Prototype B) is one waypoint within eight iterations, exploring user agency and visibility into cohort details motivated by the logic behind the data model and establishment of user trust. Prototype B highlights more data about the simulated patient's cohort. In addition to the age and sex data shown in Prototype A, it includes other factors such as other comorbidities and contraindications that could change the optimal drug choice, as well as suggesting that other insight types might be available (in this case, hypertension). Prototype B allowed the clinician to pick which factors to consider in the selection of drug regimens.

The center portion shows three best regimen options within the toggled categories selected by the user: Best Overall Insights, A1c Reduction Insight, and Fewest Medications. Since the confidence intervals as shown in Prototype A were not particularly valuable to participants, the bottom portion includes that data redesigned: an experimental interface for showing potential impact of those three treatment regimens. While the Y-axis labels are not shown in this screenshot, the fans indicate the potential range of A1c change under each possible treatment choice. Clicking on an individual insight brought up more details, e.g., an estimate of individual drug costs for each type of treatment (see Figure 3).

Prototype C (Figure 4), a screen waypoint from towards the end of the study, combines learnings from Prototype A and B while addressing the realities of its use: a quick and meaningful comparison of regimens and limited time to engage with its content. We changed to this design because some participants felt that prior visualizations gave too much detail to decipher up front, a tough ask in the time constrained environment of patient visits, and the prior design departed from their expectations of the norm of other clinical interfaces.

Prototype C includes similar information to earlier designs and adds A1c trends over time and active medications. The center portion presents a table of regimens and is the user's main decision-making area. To enable a quick read, the table starts with current

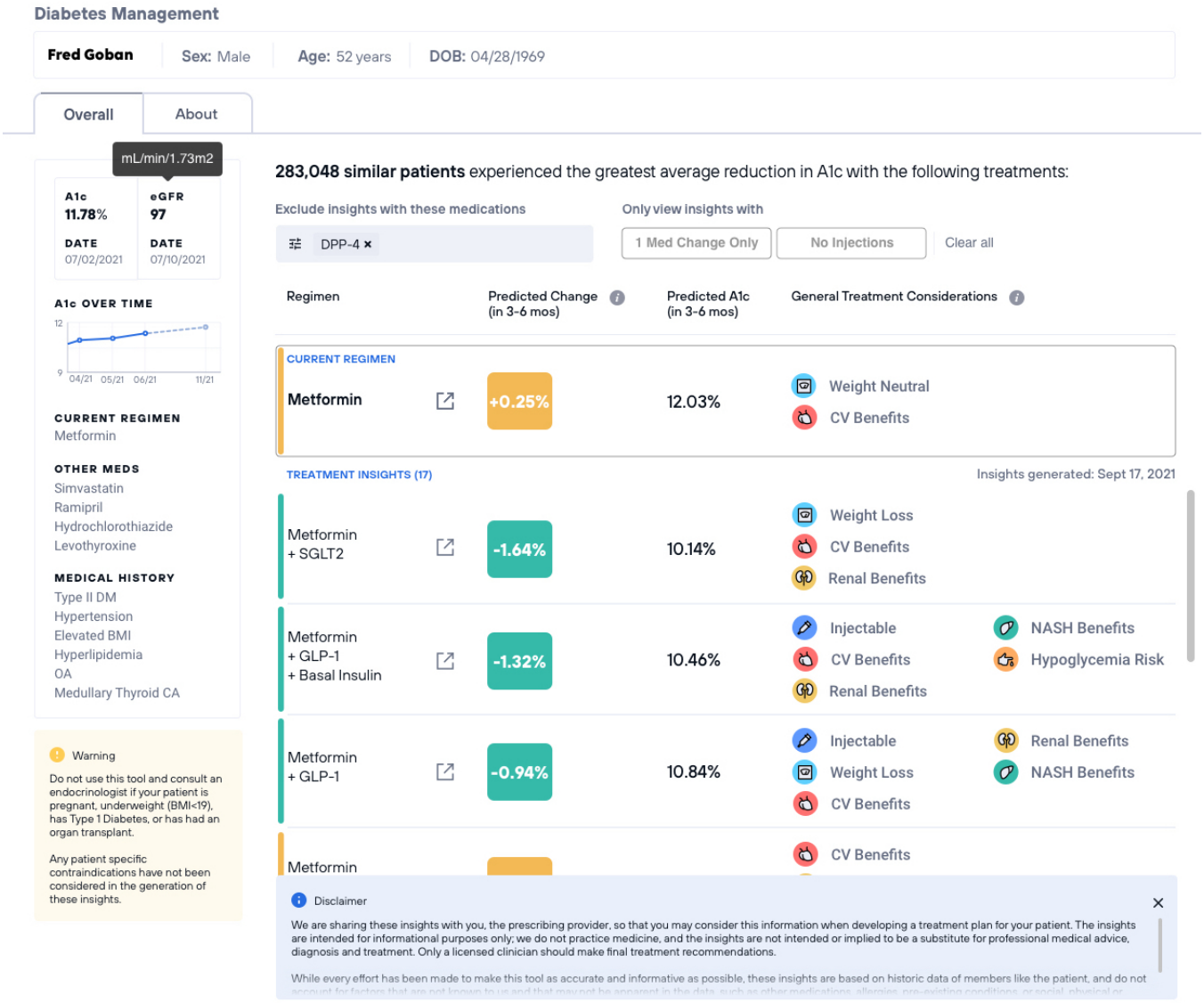

Figure 4: Prototype C shows data about a (different) simulated patient, a set of insights, and a number of treatment considerations, both positive and negative. The popup in the top left shows the standard units for measuring eGFR, mL/min/1.73m2.

regimen and contrasts it with a list of alternate regimens. The predicted change in A1c is bold and color-coded, to emphasize possible impact to the simulated patient's A1c. Given participant feedback around confidence intervals making no difference in either the treatment choices or in their trust in the underlying model, the designers did not include them and instead provided the average predicted A1c change. Additionally, this prototype allows the clinician to refine their search by excluding insights that include a particular drug class—here, DPP-4 inhibitors—and provide easy access to frequently requested filters for treatment regimens (e.g., changing only one medication from their current treatment; eliminating treatments that require injections). Several general regimen considerations were added, highlighting patient impacts (such as renal, nonalcoholic steatohepatitis (NASH), cardiovascular (CV) benefits, or weight loss), potential additional risks (such as hypoglycemia), and highlighting treatments requiring injections. Each regimen could be expanded for details: showing the medications that make up the regimen (e.g., GLP-1), their available meds in a class (e.g., Dulaglutide, Exenatide, etc.) with dosing and route, and General Safety Warnings (e.g., the need to avoid GLP-1 in case of gastroparesis).

3.2 Participants

We conducted semi-structured interviews with 41 U.S.-based healthcare practitioners (see Table 1) during a five-month period from

Table 1: Participants

Participant Notation	Role Type	Count
PCP 1–15	Primary Care Provider (MD/DO)	14
NPPA 1–18	Nurse Practitioner (NP) / Physician Assistant (PA)	18
Endo 1–5	Endocrinologist (MD/DO)	5
Pharmacist 1–2	Pharmacist	2
IM 1–2	Internal Medicine (MD/DO)	2
Total		**41**

January to May 2021. A recruiting agency selected diverse participants from practices across a variety of U.S. states, clinic sizes, and numbers of clinicians, as well as gender, years of experience, and racial and ethnic diversity. We used Reckner, a recruiting agency with a focus on providing participants in healthcare, both to recruit difficult-to-find participants like endocrinologists, as well as preserving anonymity of the process so participants were not influenced by researchers being connected to a specific healthcare institution. Our research goals were to deeply understand the workplace context and challenges of medication prescribing and to use participant feedback to carry out rounds of iteration on the AI-based prototype concept. This study was determined to be exempt from IRB review by the WCG Institutional Review Board. Participants were compensated for their time.

By interviewing a diverse panel of practitioners, we wanted to ensure that we understood the workflow processes of a variety of healthcare roles. We also sought to understand the perspectives of participants working in primary care clinics (the first and most frequent clinician many patients will see) as well as individuals working in what are commonly called "specialist" clinics (these doctors—in this case, endocrinologists—often require a referral from the primary care doctor for patients who could benefit from their specialist knowledge).

Inclusion criteria required that participants (except for the pharmacists) currently provide T2DM care and actively prescribed T2DM medication. For context, depending on a U.S. state's guidelines, not only Endocrinologists and Primary Care/Internal Medicine physicians can prescribe medication but also Nurse Practitioners and Physician Assistants as well as some Pharmacists. Therefore, all the interviewed participants had the potential to prescribe T2DM medication. Our recruiting process began with Endocrinologists, and then as stated in the Findings, we learned that Endocrinologists felt that the CDS was not as valuable to them as it could be to generalist roles. We then interviewed Primary Care and Internal Medicine physicians and Nurse Practitioners and Physician Assistants. Finally, knowing that Pharmacists can prescribe medications and critically that they often have cost-related conversations about medication with patients, we felt that there was potential relevance of pharmacist knowledge to the larger sociotechnical system of diabetes medication prescribing, so we sought to gain their feedback on the tool as well. The pharmacists interviewed in this study were involved with Case Management programs for T2DM care.

3.3 Data Collection

Three human-computer interaction researchers conducted remote, one-on-one, semi-structured interviews with the participants listed in Table 1 via Zoom. Sessions lasted between 45 minutes and one hour. The study sessions began with questions probing participant experiences of providing care to patients managing diabetes. In particular, the interviewer sought to understand current medication prescribing workflows. Then, the researcher shared their screen to present the image of the prototype tool and led a discussion. The researcher probed the participants' initial reactions to the tool and asked targeted questions about components of the prototype. Following research practices similar to previous literature [27, 31], the researcher also used the prototype to prompt further reflections on current practices and ongoing patient care challenges. As we gained insights during these interviews, we conducted dozens of iterative design revisions to align the prototype's interface and workflows with the sociotechnical complexity and clinical needs uncovered.

3.4 Analysis and Researcher Position

Two team researchers led the analysis process and regularly discussed ongoing analytical approaches and emerging themes with the research team. Our data corpus included 677 pages of transcripts. The interviews were transcribed and then analyzed following Braun and Clarke's thematic analysis method [17]. Following the steps of the thematic analysis process, we began by open-coding several participant transcripts to gain familiarity with the data. We primarily coded at the sentence level, but some coding took place at the paragraph level to understand sequences of medication prescribing and decision-making. Through a process of iterative analysis and comparison, we arrived at a set of axial codes reflected in the themes of this paper: decision optimization tensions between population-level and personalized insights, core responsibilities of the clinician, and patterns of use and trust of AI systems.

Our author team has HCI and clinical (including subspecialist diabetes) backgrounds, with many years of experience working in digital health technologies. Our team's interdisciplinary expertise was helpful to deeply understand the mindsets of our participants and to consider the important characteristics of this sociotechnical setting. To make the best possible prototype we followed a user-centered research approach to support participants in expressing their lived experience, workflows, challenges, and support opportunities.

4 FINDINGS

First, we share in Section 4.1 what it means to optimize a medication selection decision for T2DM and how those findings were reflected in prototype design iterations. Then, we share in Section 4.2 workflow findings and patterns of use of a CDS tool in the clinic, and finally describe in Section 4.3 considerations relevant to trust for AI-supported clinical decision support technologies.

4.1 "Optimizing" the decision: Distinction between 'population-level' insights ('people like you') and n-of-1 personalized medication insights

Our prototype used claims and other structured medical data as an evidence base for the AI model. The medication insights produced by the AI present comparisons with other people of similar age and health status [6]. We represented this process using text such as: "there are 4,781 patients just like Jennifer with Type 2 Diabetes" (Prototype B) and "286,058 similar patients experienced the greatest average reduction in A1c with the following treatments" (Prototype C) to illustrate that the insights stemmed from cohort comparison. However, while several participants viewed the prototype positively, especially in the comparison of certain combinations of medications, they also noted that optimizing for A1c reduction is only part of the ultimate medication selection decision. Critically, participants highlighted that T2DM medication effectiveness also depends on how well the patient can fit the medication regimen into their everyday life. Medication decisions are personalized and optimizing for A1c reduction alone does not match the prescribing decision complexity. Participants emphasized that medication insights need to balance elements of biological efficacy, affordability, and patient lifestyles and preferences. Further iterations of the protype thus included options for personalized filtering (Prototype C). See Table 2 for considerations participants described when they prescribe medications for diabetes management.

4.1.1 T2DM Medication decisions include balancing efficacy, affordability and patient lifestyle. Participants weighed several factors when recommending medication for a specific patient. Determining what treatment(s) will be the most "effective" is a complex, patient-specific question. Our initial prototype presented insights for optimized A1c control. Participants corroborated the importance of this outcome, describing how keeping down A1c is a key goal in T2DM management. For example, PCP 3 noted: "*I had somebody in yesterday. I mean they've been in good control [of their glucose] but I don't know, over the last six months their control has deteriorated not horribly from a hemoglobin A1c of 6.8 to 8. Which, I certainly try to keep folks below 7.5 if I can.*" PCP 3 and others discussed that they often request diabetic patients to have their blood drawn in time for the results to be ready for discussion during their visit. A1c readings can change over time, but it can be unclear what is influencing shifts in the score (e.g., diet, exercise, medication, other illnesses). Participants also noted some external incentives in their prescribing decision-making. For example, some participants who practiced under Value-Based Care payment models explained that the number of diabetic patients under a specific A1c reading was also monetarily incentivized in end-of-the year payments.

Participants also described that A1c control may sometimes be a secondary concern in T2D prescribing decisions. Medication costs can often guide specific regimen selection. At the moment of prescription, participants stated that they often do not have the complete details about the final cost of a medication for a patient. When reviewing expected cost shown within the prototype, Pharmacist 2 asked if the tool drew on information about the patient's specific insurance plan and their deductibles, which may change over the course of a calendar year:

> "*Would it reach down as far as seeing if that patient has met their [insurance] deductibles for the year? So is that just a block price that [the prototype] has pulled out of some database? [Do] they get specific as to that patient's exact cost, and that exact cost, as it progresses throughout the year? Because not necessarily what a patient pays in January is what they're going to be paying in August.*" (Pharmacist 2)

In the United States, there are many healthcare insurance companies, including some government-supported insurance options (e.g., Medicare and Medicaid). Participants described how each insurance plan may negotiate and determine coverage for medications individually, so one patient's healthcare insurance may cover a particular medication, and another patient's healthcare insurance may not. In addition, some name-brand drugs also have patient assistance programs. Furthermore, medical plans often have deductible limits that need to be met before the plan starts paying for medications or medical services: for example, the patient pays the first $500 of any medical costs in a year. After the deductible is met, some insurance providers will cover part of the cost of a medication, leaving the patient to cover the remaining amount. The amount a patient pays is called their out-of-pocket cost(s).

Ultimately, the variability of patients' healthcare insurance coverage(s) (as well as lack thereof) introduces strong uncertainty into the actual cost of a patient's medication when they pick it up from the pharmacy or have it delivered to their home. PCP 2 described how lack of insurance coverage led to a recent patient "*haphazardly picking his medicines because he couldn't afford to get 'em.*" Similarly, PCP 13 noted that throughout the COVID-19 pandemic, there were many layoffs, which could eliminate a patient's employer-sponsored health insurance, requiring clinicians to go "*back to the drawing board*" (PCP 13) for more affordable medications. Therefore, an individual's ability to pay out-of-pocket and the exact coverage of their healthcare insurance can influence which medications are ultimately prescribed. Accurate patient-level costs are thus a desired feature of a CDS tool for diabetes management.

Next, taking medications creates burden in addition to other life responsibilities. To reduce patient burden, and sometimes to also reduce cost, fewest medications was another important consideration. NPPA 11 describes how a patient request for the fewest number of medications influences her regimen recommendations:

> "*I'm like, yeah, some of these [medicine insights] make really reasonable sense and they're game-changers in the management of diabetes. But if they [patients] say least medications, then you know, you at least can take them off your radar and just think about the other choices quickly*" (NPPA 11).

We incorporated these findings into Prototype B with an optimization toggle (see Figure 5) where clinicians could optimize based on several factors including A1c reduction, lowest cost treatment, and/or fewest medications.

Later versions of the prototype (e.g., Prototype C) were designed to support "shared decision making" [43], where clinicians and patients could review treatment options together and tailor regimens

Table 2: Type Two Diabetes Mellitus Medication Decision-Making Considerations

Consideration	Rationale	Representative Quote
HbA1c Reduction/ Glucose Optimization	Better blood sugars control can help reduce long-term effects of high blood sugar including kidney failure, blindness, heart attacks, strokes, infections, and amputations.	"*I certainly try to keep folks below [hbA1c] 7.5 if I can*" (PCP 3)
Medication Costs	A patient's ability to afford treatment is based on their specific health insurance coverage and ability to afford out-of-pocket medication.	"*Or you know 'at pharmacy the medication was too expensive. I didn't pick it up and I didn't think to call you.' I've had plenty of those. 'Oh the pharmacy said I needed [prior authorization] and they were supposed to message you but they never did.'*" (NPPA 18)
Number of Medications	The number of medications that a patient will take has implications for complexity of purchasing (e.g., at the pharmacy) and complexity of taking medications (e.g., many pills per day). Similarly, this is an important consideration when an individual is also taking medication for other conditions (polypharmacy).	"*But if they [patients] say least medications, then you know, you at least can take them [some medications] off your radar and just think about the other choices quickly.*" (NPPA 11)
Drug Delivery and Storage Method	Some diabetes treatments require injection. When patients are worried about or unable to use needles, clinicians may lean toward oral medications. Also includes requirements for drug storage and dosing complexity.	"*If he is adamant to not add any more injectables, I could talk to him about some of the oral medications. We may not get as much efficacy as another injectable, but that's something we could try.*" (PCP 2)
Patient's Daily Routine	Includes questions about the requirements of the patient's daily life including their job, childcare responsibilities, location, and number of breaks day-to-day, and many other aspects with relation to the ability to carry out certain medication regimens.	"*What they do work-wise, how hectic their schedule is, you know, are they going to be able to take something, um, 30 minutes before? Are they like rushing out the door in the morning? And then when it comes to once a week or every day just again their lifestyle.*" (NPPA 2)
Secondary Benefits	Some diabetes medications have beneficial impacts for other organ systems, e.g., renal (kidney system) benefits, cardiovascular benefits. Some medications also support weight loss which may also be beneficial (e.g., for mobility increase).	"*Both of those classes will typically help with weight loss and neither, neither one will cause hypoglycemia. So they tend to, and they both offer a cardiovascular protection.*" (Endo 3)
Potential Contraindications and Side Effects	Risks and side effects associated with certain medications (e.g., going hypoglycemic, introducing a fall risk for elderly patients, adequate kidney function).	"*He's had no hypoglycemic events or severe interactions and he has no symptoms of any kind of renal impairment*" (NPPA 1)

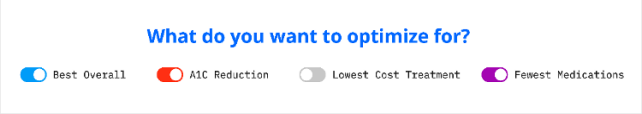

Figure 5: Redesigned Toggle for Optimization Selection (Prototype B)

with patient preferences (such as avoiding injections) considered (see Figure 4). Such changes improved clinical relevance. For example, regarding drug delivery and storage method, participants described how some patients were frightened of needles. Therefore, even if an injectable medication was the recommended clinical treatment, physicians were likely to recommend an oral medication instead. "*If he is adamant to not add any more injectables, I could talk to him about some of the oral medications. We may not get as much efficacy as another injectable, but that's something we could try*" (PCP 2). Given that this was a key medicine characteristic that is important to discuss with a patient, we highlighted this in the prototype (see Figure 4) with a "Injectable" callout. Participants also considered how the medication needs to be stored and whether the patient has the ability to do so (e.g., having access to a refrigerator for certain medications). Relatedly, dosing complexity—how many times a person needs to take the medication over the course of a day—was also an important factor. For instance, metformin has two formulations, one taken once and the other taken twice a

day. Other medications, such as the GLP-1 class, are injections but taken weekly, whereas insulin is usually a once- to four-times-a-day injection or infused by a pump.

Drug delivery method also has implications regarding how it fits into the patient's daily routine:

> "*I kind of take into consideration, you know, their life, you know what they do work-wise, how hectic their schedule is, you know, are they going to be able to take something, um, 30 minutes before? Are they like rushing out the door in the morning? And then when it comes to once a week or every day just again their lifestyle.*" (NPPA 2).

While we did not create a daily routine call-out in the interface, routine is an essential backdrop to medication decision-making.

Some diabetes medications have secondary benefits. For example, beneficial impacts for other organ systems such as the renal (kidney) system or the cardiovascular system. Some medications also support weight loss which may also be beneficial (e.g., for mobility increase, or decreased insulin resistance). When discussing medicine selection, Endo 4 described how she will often prescribe two main first-line treatments, saying "*both of those classes will typically help with weight loss and neither, neither one will cause hypoglycemia. So they tend to, and they both offer a cardiovascular protection.*"

Endo 4's quote also shows a consideration for the potential contraindication of hypoglycemia. Potential contraindications and side effects may be associated with certain medications (e.g., hypoglycemia, fall risk for elderly patients, inadequate kidney function). NPPA 1 notes that he watches out for his patient's potential contraindications, saying about a specific patient: "*he's had no hypoglycemic events or severe interactions and he has no symptoms of any kind of renal impairment*," whereas NPPA 18 describes a patient she prescribed medication to and then "*they started taking it and they got side effects*" (NPPA 18). PCP 2 notes that he sometimes struggles to remember potential contraindications: "*there's so many out there that it's difficult for us to always remember all of 'em and maybe having something like this [the prototype] would be very helpful*" (PCP 2).

In summary, algorithms that over-optimize on outcomes metrics can lead to unrealistic insights for real world clinical decision making. The clinician needs to understand what is being "optimized" in the insight so that they can interpret the information and add it to their larger decision-making process. There is an opportunity to design AI-insight processes to consider patient-specific factors that clinicians weigh as they make medication recommendation decisions, as we have done in our prototype (see Figure 2, 3, and 4) to better match the complex interrelated factors important for these decisions.

4.2 Core Responsibilities of the Clinician

Another factor relevant to decision support is how the system insights sit in complement to the core responsibilities of a clinician regarding their expertise and the requirements of the patient visit.

4.2.1 Perceived Expertise. A tool that shares insights that a physician feels they already know well is not likely to be seen as valuable. For instance, some participants looked at insights generated from the prototype, commenting, "*the big picture is I don't turn to an EMR to tell me how to treat a simple – not simple, but… – a disease like diabetes, which is something that I treat on a daily basis. I use 20 years of clinical experience*" (IM 2) or "*I know what the options are for treating diabetes*" (IM 1). T2DM is a common diagnosis in the United States, and therefore, many physicians have frequent experience diagnosing and recommending treatment and lifestyle changes relevant to diabetic patients. All participant endocrinologists felt that they did not need a tool to recommend medications for diabetes because this is their specialty area. One endocrinologist participant said regarding the insight calculations: "*most experienced doctors are probably doing that in their head without the help of this*" (Endo 3). However, other participants including a few primary care doctors, nurse practitioners, and physician assistants, noted that they thought it could be a useful tool to compare against their line of thinking and surface potential contraindications. These comments clarified for the design team that individuals with less specialty training may be more interested in this type of CDS tool.

Other participants noted that what might be simple for a clinician to understand could be useful to show a patient. For example, reacting to our prototype: "*Oh yeah, that's good...You could print it for the patient*" (NPPA 17), which could support shared decision-making regarding the complex considerations surrounding medication regimen choice. NPPA 3 agreed with this sentiment, noting that "*I think that could be a joint decision-making tool*" (NPPA 3), but stressed that she thought it would be most useful for discussion with patients on complex medication regimens.

4.2.2 Patient Visit Time Constraints. The limited time available for the clinical encounter is an essential design consideration factor. Participants consistently described the time-constrained nature of a patient visit. Participants focused effort on having face-to-face time with the patient as much as possible instead of looking at a computer screen. Face-to-face communication is important to facilitate connection, information-sharing, and open discussion. Particularly for T2DM, lifestyle changes and medications can be challenging to self-manage and require significant patient effort and understanding. Therefore, our participants described the importance of getting patient buy-in for these processes to manage their diabetes. However, face-to-face communication can be challenging when much of a clinician's workflow includes documenting notes and e-prescribing conducted via digital device (e.g., desktop computer, laptop, tablet). While the prototype could potentially provide a substantial amount of detail regarding future predictions, it was clear that a patient visit is not a time when clinicians can do research. "*I don't know, per se, that I have the time during each visit to kind of go into this*" (Endo 3). Therefore, insights should be delivered to support the time constrained needs of clinicians in the moment.

In addition to summarized insights, clinicians described moments in a patient's treatment when a medication insight tool would be most helpful. When a previous medication was not as successful as desired, clinicians were most interested in next-step treatment insights from a CDS tool. Primary care participants described that this is often the point when they refer patients to endocrinologists. NPPA 2 described how she would seek the advice of the prototype when a need was identified to change her patient's medications: "*If I*

was going to change prescriptions for a patient. So after we determined 'hey a change is necessary', I would probably use it there" (NPPA 2). This quote underscores that there are certain times in a patient's illness trajectory when medication insight tools become especially helpful. This learning prompts further design questions: should the prototype show insights for patients with an A1c $\geq$ 9% as part of the workflow for every patient visit, or are there are certain triggers (e.g., threshold of A1c percentage increase) that could "push" the insights into a clinician's workflow?

Overall, many participants felt that they had a good grasp of the basics of T2DM medication prescribing. Yet many participants also had patients who were above the desired A1c threshold. Participants stated that gaining patient buy-in to take the prescribed medications and factors outside of specific medications such as diet and exercise played a major role in the scores. However, clinicians also need to be aware of new medications and their effects. Indeed, IM 2 notes this, saying: "*Most patients need two or three different products and there's different combinations of products and there's always new things coming out. So I try to stay abreast of the new information and see what I can apply to the patients*" (IM 2). A CDS system which pulls from recent medical information may be able support ongoing learning about new medications. Finally, some patients may prefer to optimize for the very lowest A1c score possible, and therefore having a view which optimizes for elements which are most important to patients may better support shared decision-making beyond current approaches.

4.3 Patterns of Use and Trust in AI Systems

In this section, we discuss findings regarding what we call the "front-loading" of trust for AI-informed CDS and trust related to the type of data powering the AI models.

4.3.1 "Front-loading" trust for AI-informed CDS. Participants reported wanting to determine their trust of an AI insight system when first introduced to the tool. When participants saw an AI-supported insight for a medication or selection of medications, they wanted to know how the insight was calculated and the outcomes the insight prioritized. For instance: "*And where they're pulling this information from, is it from a EHR? Is it thorough? Up-to-date?*" (PCP 2) and "*I'd like to see the data and I'd like to see, you know, who made it*" (NPPA 3). Once they determine that trust, they can quickly utilize the insight within their workflow. Ultimately, technologies used at point-of-care need to be streamlined to ensure that a clinician can spend minimal time clicking through screens and can focus their attention on the patient. Our participants did not want to validate insights every time they used the CDS tool. Unless the insight was unexpected, in which case participants wanted to look more closely at the underlying logic (similar to findings from Jacobs et al. [31]), participants did not expect to spend time investigating the underlying rationale behind each AI-produced insight. Instead, they will determine trust once, based on initial use experiences and contextual credibility (e.g., the backing of medical institutions). In short, participants underscored that the care visit is not the right time to do deep research. The onus for trust-building rests on the introduction of the tool and during the first use or first few uses of the technology. Relatedly, some clinicians may be more personally open to AI systems. For example, upon learning that the CDS was powered by AI, NPPA 11 noted: "*I'm welcome to learn anything and I'm welcome to look at trends and I'm welcome to see how that would affect my patient and my prescribing*" (NPPA 11). We did not test the marketing and onboarding of AI-supported CDS tools, but we believe these first exposures to an AI-powered tool are important future avenues to investigate for building trust in AI-powered tools.

4.3.2 Trust in the organization(s) and data powering AI models. For healthcare technologies, clinicians often consider the potential incentives of the technology's organization. For example, participants described discomfort trusting a CDS tool for medication insights if it was sponsored by pharmaceutical companies. However, if an insurance company were providing the tool, NPPA 2 described that she viewed the system similarly to a formulary (lists of medications covered by an insurance plan) viewed as canonical sources of information. These perceptions demonstrate the importance of clearly identifying the organization that developed the model as well as understanding associated user expectations (e.g., the expectation that medications recommended by an insurance company be covered for a particular patient).

Participants also discussed trust in the data itself. Our prototype used claims data in addition to labs and medication fill records. Some participants described previous experiences seeing the outputs of claims data through other reports and technologies. Unfortunately, some participants noted that these tools were often out of date or inaccurate, whereas others felt that claims data was helpful. For instance, "*[Claims data] seems like the one that lags behind the most, I mean, it seems like the most erroneous or inaccurate data thus far that we see*" (PCP 2) contrasts against PCP 1's view:

> "*I think insurance claims data is very accurate. Most of the time, you know, they're putting together hemoglobin A1cs from thousands of patients across the nation. You know, I trust their data. They trust their data too. Cause they're making big money decisions based on that data.*" (PCP 1)

Similarly, other participants talked about data connectivity experiences which led them to be more excited about an AI-support CDS tool. Pharmacist 2 noted:

> "*I was a member of task force [for] our state to interconnect hospitals and labs and doctor's offices. And we actually got multimillion dollar grants from the federal government to do that. And it's a very similar concept pulling in data from the global archives of that patient's record rather than just maybe an individual patient record within our specific institution.*" (Pharmacist 2)

Overall, previous experiences with tools built on certain types of data are important context for technologists to understand, particularly when those experiences have been negative. Even if the validity of the AI insights is shown and documented in peer-reviewed research, if claims data in general is viewed negatively by providers, then it is more difficult to build trust in a tool which pulls from these databases. Our model ameliorated this somewhat by also pulling from labs and medication fill records in addition to claims data.

5 DISCUSSION

Our findings present the rich sociotechnical complexity of using a CDS tool within the workflow of a clinical visit and critical facets of trust when presenting an AI-based clinical decision support technology. We discuss in Section 5.1 provider perspectives on AI as a new form of knowledge production and in Section 5.2 six design principles for AI-supported CDS systems.

5.1 Healthcare provider perspectives on AI as a new form of knowledge production

As Vereschak et al. [61] state, building a "collaborative partnership between human deciders and AI-embedded systems…critically relies on trust from the users toward the systems." Our paper's findings point to various factors affecting trust in CDS systems. For example, our participants looked for validation of medication recommendations from two places: guidelines put out by professional societies and papers published in (reputable) clinical journals. Guidelines are seen as more solid recommendations, but clinicians also recognize that the guidelines are slower to change than findings from publications. For example, the American Heart Association updates their Professional Standards of Care annually [3]. However, these guidelines do not necessarily reflect the latest research: questions around the role of cardiac health in Long Covid-19, for example, are being addressed in the medical literature [52], but more than two years after the start of the pandemic they have not yet made their way to the status of fully endorsed guidelines.

At a fundamental level, trust is vested in the underlying mechanisms of generating knowledge in clinical sittings, notably the randomized clinical trial. The randomized clinical trial process itself is not without controversy [21] and has long been criticized for its sometimes arguably weak relationship to the lived reality of medical experience. Candidates signing up for a clinical trial have been shown to be disproportionately more male, younger, and sicker than the population who would use the clinical intervention or drug in question [35]. Many clinical trials exclude individuals with multiple conditions (e.g., both diabetes and cancer) [10]. Furthermore, as clinical trials are often funded by the organization that has developed the intervention or drug in question, be it corporate or academic, their incentive is to show the efficacy and safety of that drug, and not necessarily the effectiveness of that drug as it would be prescribed and taken in conjunction with other drugs. In the context of these limitations this work speaks to the clinical area known as real-world evidence (RWE), based on analysis of real-world data (RWD). RWE is "the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. RWE can be generated by different study designs or analyses, including, but not limited to, randomized trials, including large simple trials, pragmatic trials, and observational studies (prospective and/or retrospective)"[23]. The benefit of RWE is that it provides insights into the use and outcomes of treatment in everyday practice.

However, data need to be available in the first place for systems to use that information. For instance, Wang et al. [63] noted that in rural clinics in China, there is a major shortage of medical staff, especially nurses who check blood pressure, record medical history, take a person's temperature, and more. Thus, for their CDS, the authors note that "it is impossible for the clinicians to capture sufficient information for AI-CDSS to make accurate and comprehensive diagnosis." This deficit means that, depending on the purpose of an AI-supported CDS, the level of personalized insights the system is expected to support (e.g., diagnosis of an individual vs. recommendations for medication for 'people like you') will necessitate different types of required work to input new patient-specific data over time.

We also found that providers' individual perspectives regarding the data underlying the models could influence the trust around the outputs of an AI-supported CDS. This finding resonates with recent literature calling for additional research to understand the role of individual differences in trust such as prior experience and self-confidence [61], general familiarity with AI/ML [32], and the relationship between a person's faith in and perception of the capability of automation [46]. Jacobs et al. [32] found that clinicians with higher familiarity with ML were less likely to use an ML recommendation compared to clinicians with lower ML familiarity. The variety of these trust-relevant topics point to the importance of individual perceptions that may differ from clinician to clinician. Thus, we recommend that strategies for developing appropriate levels of trust in these models should be in part tailored to individual users.

We join many other authors in this space (e.g., [13, 32, 46, 61]) to argue that as a new tool for knowledge discovery, AI-supported CDS has an onus to build trust and impart some element of the decision-making process to clinicians. The black box nature of many previous and current AI systems does not easily engender trust, and, in this relational sociotechnical experience between clinician, CDS, and patient, a black box output does not provide a clinician with persuasive rationale to communicate recommendations to patients. Beyond better explanations, researchers, developers, and designers of these tools also need to understand clinician needs and support their efforts to integrate AI insights as part of their overall decision-making process. Rarely will an AI-supported tool be a clinician's only decision support resource. Rather, they mix different forms of knowledge. We need to consider how to support a clinician making sense across multiple resources, for instance across recent journal articles, AI tool(s), and other decision support resources (e.g., norms within a clinic/hospital system; printed resources on the wall; rules-based alerts). Specifically, many recent articles have positioned AI tools in the singular, for instance stating that a clinician must understand "the" AI system in relation to their own knowledge and learn its capabilities. Yet it is likely that there will be many AI-supported systems in the future, and so the complexity of understanding each system as the technology ecosystem grows is something to consider as we build future mental models and training.

We must also acknowledge key fears discussed in the broader literature regarding human-AI collaboration vs. concerns of replacement (e.g., [29, 55, 62]). The intention for our prototype is to augment a clinician's knowledge, not as a replacement. However, as recent CHI research shows, AI tools have begun to be integrated into many decision-making contexts with high-stakes outcomes including human resource management [50] and child welfare work [34]. Given what we see in these domains, at some future point, clinicians may be required to refer to AI-supported

insights as part of their day-to-day work. We are at an early stage of the likely long future trajectory of AI tools in medicine. Now is the time to get the basics right to ensure future tools are developed with an understanding of the broad sociotechnical complexities of patient encounters and specific prescribing complexities within clinician-patient conversations.

5.2 Design Principles for AI-Supported Clinical Decision Support Systems

We present 6 design principles for development of this T2DM AI-supported CDS system and similar tools:

(1) Account for what is possible and realistic for the patient and the clinical context. Algorithms that over-optimize on disease outcome metrics can lead to unrealistic insights.
(2) Give the clinician the ability to weigh patient-specific factors that cannot be easily inferred automatically; give the clinician agency/control over model output.
(3) Do not introduce "research" tasks for clinicians into patient visit workflow.
(4) The introduction of the AI tool is a core opportunity for trust building.
(5) Create networked systems designed for collaborative use by patients and healthcare staff throughout the patient's care pathways.
(6) Pinpoint where complex decisions need to take place in a clinical workflow versus tools that provide blanket data that physicians already know.

5.2.1 Account for what is possible and realistic for the patient and the clinic context. Our research highlights the reality that algorithms that over-optimize on disease outcome metrics can lead to unrealistic insights. Medication insights need to balance elements of biological efficacy, affordability, and patient lifestyles and preferences. For example, field workers may be unable to keep insulin at cold enough temperatures at work, and cardiothoracic surgeons may want to avoid medications that increase urinary frequency. We discussed the many considerations guiding medication selection in Section 4.1, including being able to adapt CDS insights and workflows based on a clinic's contextual constraints, as described by several recent studies (e.g., [4, 63]). For instance, Wang et al. [63] noted that rural Chinese clinics have very limited medications in stock and that their studied AI-CDS suggested treatment options that clinicians could not prescribe. Clearly both clinicians and AI-CDS sit between questions of what might be most optimal medically and what is feasible realistically given contextual factors. Thus, we must consider how to "personalize" insights to the specific local context of the clinical settings. We add to the literature by presenting a multi-design objective rationale through which filters enable personalizing CDS insights to the patient.

5.2.2 Give the clinician the ability to weigh patient-specific factors that cannot be easily inferred automatically; give the clinician agency/control over model output. As we highlighted with our prototype iterations, there are many consideration factors relevant to a clinician's personalized medication recommendation for T2DM management. AI CDS tools can help support that personalization process. Critically, the clinician needs to know what the algorithm is optimized for (the design objective [13]). When the design objective is clear, we contribute to the HCI literature by showing how personalization can be refined. In the context of our CDS tool that refining process operates through filters in the tool and through shared decision-making between clinicians and patients and their caregivers. As described in our Findings, medications are most useful when they fit into a patient's life. No matter how potentially helpful a medication might be, if a patient cannot afford it or they are unable to incorporate it into their daily routines, then it is not a reasonable choice for them. We envision tools like our prototype allowing a clinician to review potential future states with a patient and discuss options while viewing the tool together, e.g., "I know you don't like injections, so we can achieve a similar improvement in your A1c with this combination of oral medications" or "unfortunately, without an injectable medication, it's unlikely we can get your A1c to goal."

5.2.3 Do not introduce research tasks for clinicians into patient visit workflow. As our findings and other related research [4, 63] show, the patient encounter is not the best time for recommendations that require time to evaluate. We described how our participants were both unable to and uninterested in validating an AI-CDS's insights every time they used the prototype. This finding then prompts the question: when and how might we most effectively surface CDS insights? In their paper [60] observing providers using EHRs for diabetic prescribing at the Veteran's Association in the United States, Veinot et al. noted clinicians' concerns that EHRs did not provide a succinct summary of diabetes-relevant information all in one place. Their clinician participants wanted to see this information during the "priming" process where they read details about the patient during the planning period before an upcoming visit. To address this gap, perhaps the pre-visit review is an opportune time to highlight AI-supported medication insights if there is a suggested improved medication combination for a particular patient, although it may diminish the opportunity for shared decision making.

5.2.4 The introduction of the AI tool is a core opportunity for trust building. Given the current time constraints of patient visits, clinicians simply do not have the luxury to investigate and build trust over time. Similarly, lack of training can mean that users are not aware of useful features and may worsen transparency and trust issues [63]. Therefore, the introduction and first uses of an AI CDS are critical to get right. Clinician users will likely pay close attention to both the initial user experience and contextual credibility as part of their trust determination process. Cai et al. [14] have underscored the importance of onboarding materials as key opportunities for supporting human-AI collaboration and engendering trust. Connecting our findings to potential future onboarding materials as we noted in Section 4.3, clinicians carried over their experiences with other tools, such as those that use certain types of data (e.g., claims data), as part of their initial reaction to any type of tool which used those same data types. Therefore, we extend lines of thinking in related literature to recommend that, in addition to sharing the design objective(s) of the AI-supported CDS, there should also be specific discussion of the data, particularly its acquisition process, underlying the model as part of onboarding materials.

5.2.5 Create networked systems designed for collaborative use by patients and healthcare staff throughout the patient's care pathways. Clinicians have a rich set of activities that require interfacing with a patient's medical record and history over time. While much prior work focuses on a single decision support period, revising prior clinical decisions, sharing information across often fragmented healthcare providers (both fragmented informationally with different EHR systems, and organizationally through different specialties and focus areas), and sharing collaborative decision-making discussions regarding an ever-changing health future all require attention to the temporal elements of a person's illness trajectory. The prototype we tested in our study was designed to support an in-the-moment decision by a certain healthcare practitioner in consultation with a patient. However, we recognize that patients may have many clinical encounters along their illness trajectories. Patients may see many different clinicians and may be diagnosed with other comorbid conditions over time. A decision made one year by a primary care physician might need to be revisited by a specialist physician a couple of years down the road, or vice versa. Therefore, we must design CDS tools with this larger care team and a networked approach in mind. Relatedly, extending this temporal lens to the long-term use of the tool itself, AI-supported CDS will also need to show accurate updated information when guidelines on how to take medications shift over time, necessitating a connection to ongoing research and knowledge.

While the prototype we investigated provided first-step visualization about potential future illness trajectory biometrics, we believe there is more needed work to support visualization of potential differences between treatment decisions to facilitate clinician and patient discussion. As Burgess et al. [12] have noted, when patients managing a chronic disease think about treatment decisions, they often envision a "future normal" of what their life will look like down the road given their choice of treatment. The prototype evaluated in this paper was designed with a clinician as the primary user and A1c reduction over time as the primary visualization. As we consider future design that may incorporate a patient as another primary user, particularly within shared decision-making [25] processes, there are other ways we might represent future trajectories. For instance, a lowered A1c level might allow a person to achieve future goals to be present for a grandchild's graduation or other future states they look forward to being able to do.

5.2.6 Pinpoint where complex decisions need to take place versus tools that provide blanket data that physicians already know. In our study, physicians noted that they did not usually struggle with which diabetes medications to recommend. There is a limited set of diabetes medications and if medications need to be changed, many patients come in already having tried a few medications previously, reducing the pool of potential future medication candidates. However, the fact that the prototype tool is aimed towards patients with poorly-controlled T2DM indicates that, so far, effective treatment regimens have not yet been tried for these patients. Additionally, there are opportunities for other AI-supported CDS medication prescribing technologies such as providing insights regarding experimental cancer medications which are constantly at the cutting-edge of research, or as Jacobs et al. [31] advocate for: treatment insights for depression when the first-line drugs (medications that are commonly recommended first to patients) fail to be effective.

When there are many possible medication options, having a CDS tool to help organize and make sense of the possibilities is helpful. When the next-step choices are more constrained, whether that be through cost and insurance coverage, the particular disease, or other aspects, a CDS tool may be viewed as less helpful. One way to approach this line of thinking is to consider the boundaries of a clinician's knowledge. For example, PCPs often see diabetic patients and commonly prescribe both medication and recommend lifestyle shifts to support A1c reduction. However, when their first-line process is not successful, those second- and third-line medication insights often have more options that a clinician may be less familiar with [56] and therefore the CDS can provide more perceived value. Relatedly, Wang et al. [63] found that clinicians liked and used a "similar case" feature allowing them to look up similar patient cases from top-tier research hospitals, enabling self-educating and learning in context. In contrast, an endocrinologist will be more familiar with third- and fourth- line options, and, as an expert in the field, will often be able to adopt new medications based on clinical trials faster than an AI tool can learn from retrospective patients, thus the CDS tool will likely have less perceived value for specialists. Similarly, a focus of our prototype was showing recommended combinations of medications, for which there is less clinical trial evidence. There are undoubtedly more domains where the benefits of AI can harmonize with the expertise of clinicians, and we believe this should be a key focus area for future human-AI trust research.

6 LIMITATIONS

This study presents the views of clinicians in the United States. We noted U.S.-specific findings such as lack of clarity around the cost of healthcare services (given the nation's particular insurance and pharmacy cost coverage elements). However, the central findings of our paper regarding trust of AI results and comparison against clinical trial findings are broadly applicable given the history of medical training and how medical best practices are determined in many other countries.

Second, this study introduced the prototype tool to clinicians in the context of an interview study. This supported prototype development through rapid iterations to match the needs of clinicians in a patient visit. However, participant insights represent prospective use gathered from a non-interactive prototype. Insights about trust and confidence are based on participants' impressions of seeing the tool without interacting with real world data; therefore, validation studies involving interactive and functional prototypes are needed to better understand the potential influence they have on clinicians' trust and confidence. To continue to explore how this tool would be used in the clinic, the next step is to deploy a pilot to further develop our understanding of its use.

Third, the prototype evaluated in this study is disease specific to T2DM and the model was only trained on patients with an A1c $\geq$ 9%. The prototype did not combine insights across disease states (e.g., if a person has both diabetes and chronic kidney disease). As noted in our findings, medication considerations are impacted

by comorbid conditions and thus these interrelated aspects are important considerations for future design.

Fourth, in this study we analyzed the responses of 41 participants. However, the collective framing may leave room for evaluated nuances between different types of clinicians. For example, nurse practitioners in our study saw the prototype being more valuable for their day-to-day work than endocrinologists, which suggests opportunities for differentiating interfaces and tools for explicit subsets of the eventual clinical user population.

7 FUTURE WORK

Areas for future work include the medical-legal issues of following AI-based CDS, e.g., whose fault is it if the AI is wrong? Participants said they would be reluctant to make aggressive treatment decisions that deviate from clinical guidelines. However, best practices to support clinician interpretation of conflicting insights remain to be determined (e.g., what if two AI models give different results?). Jacobs et al.'s [32] study of medication selection for patients managing depression demonstrates how AI/ML errors may negatively influence clinical decisions and, specifically, that explanations did not effectively address accuracy issues. The authors note that "in addition to commonly discussed issues of technical readiness and data bias, the interface design of ML decision support tools must be systematically evaluated." Participants mentioned the possibility of litigation for medical malpractice as a driving factor behind some of their decisions. As new tools are developed and become required by certain healthcare systems, new medical malpractice laws will likely emerge.

We also considered how to represent data and options in meaningful units for clinicians. Tensions arose between the representations that were meaningful from the point of view of the data, and representations that were more aligned with clinician's views of the world. For example, endocrinologists working with the data team had suggested two categories to divide up age cohorts, over and under 65 years. The data confirmed that there were no substantial differences between a cohort defined as "<65" and subsets of that cohort, such as "55-64" and "45-54". However, when endocrinologists observed the prototype in the field, they were concerned that insights for "<65" were too broad. Should the insights be split into subsets to potentially be more trustworthy for the clinician using the tool and confident that that insight was valid, or does that imply more accuracy than was initially proposed?

This study did not include patient interviews, a population relevant to this tool, which we plan to conduct in the future. For instance, should patient-specific information be sent in advance of a clinical encounter, for example, a notification that a medication discussion will happen? This might allow appropriate time for prereading if desired. And, if a clinician moves the screen in which they are viewing the EHR to show the patient, how does this experience unfold and how can the visualization best support this work? Similarly, as we consider explainability and interactivity from the clinicians' perspective, we should consider these factors from the patients' perspective as well. In Mitchell et al.'s [24] ML-powered nutrition goal recommendation app, the authors suggest offering a chatbot functionality to introduce concepts, answer user questions, and more fully explain goal recommendations. They also note that "a more interactive and conversational interaction style would also offer another approach to address the challenges of context…to allow participants to have input on their goals and negotiate." Depending on the breadth of an AI-supported CDS, we might consider having the patient be able to share insights about how well the medication regimen fits into their everyday life through interaction with, for example, a smartphone app, as one potential idea to better surface contextual self-management concerns throughout a person's illness trajectory.

A technical question from this research is how to move from population-level datasets to personalized prescribing insights that appropriately reflect sociotechnical complexities. We believe that this question is important and transferrable to other contexts. There are multiple pieces to an AI-supported CDS. For example, the prototype investigated in this study has underlying data, an AI model trained on a sample of those data, and a software system incorporating the AI model's output into an interface. Thus, there is a critical distinction between the AI model and the software product. While this is a familiar story in human-computer interaction, rhetoric focused around the model itself may encourage conflation of the AI model with its CDS tool, and thus a focus on optimization of the AI model at the expense of implementation, a tension we believe is a significant domain for future work.

When moving from a data-driven model to personalized insights, there are two general approaches. The first is to integrate the important sociotechnical components discovered in research (e.g., lifestyle factors, patient preferences, cost) directly into the model itself. The second is to use the same population-level model, but integrate a secondary stage in the product where user characteristics are inputted (e.g., via filters) to create more personalized insights. The first approach, when focused on a specific output like A1c, may have a cleaner output and be possibly more interpretable. Then, additional elements can be integrated using a simple rules-based filter (e.g., fear of needles). This work did not evaluate the second process – integrating sociotechnical components directly into the model itself. For instance, we might imagine a model which integrates weighted outcomes of medication on different organ systems. However, such a model then raises questions about how the weighting process is conducted and what the output means regarding the eventual care recommendation. Thus, how best to account for issues that are inherently situational (e.g., personalization) is an intriguing area for future research within the context of AI-supported insights reflected in prototype design iterations.

8 CONCLUSION

In this paper, we explored the following questions: How do clinicians make sense of AI-created insights in relationship to other treatment recommendation information? Moreover, how do they develop trust in these insights, given what they know about other ways of creating knowledge in the world, for instance, through a clinical trial? We used rapidly iterated conceptual prototypes based on AI-generated treatment insights for Type 2 diabetes mellitus to elicit feedback from 41 U.S.-based healthcare staff, including primary care, internal medicine and endocrinologist physicians, nurse practitioners, physician assistants, and pharmacists. Critically, we found that AI systems are judged against "gold standard" previous

methods of clinical knowledge generation, in particular, randomized controlled clinical trials. We discuss how confidence or trust in the insights may be influenced by the providers' understanding and confidence in the methods used to generate the insight, a major focus of explainable AI work. We describe findings regarding decision optimization tensions between population-level and personalized insights, and patterns of use and trust of AI systems. We discuss healthcare provider perspectives on AI as a new form of knowledge production and conclude with six design principles for AI-supported CDS systems.

ACKNOWLEDGMENTS

We thank our participants for sharing their considerable expertise and their thoughtful feedback. We also received valuable feedback and input on the prototypes and our processes from our colleagues: thank you, in particular, to Ashley Dumm, Henrik Olsen, Andrew Bellay, Kurt Iverson, Niki Panich, Alex Ruggieri, Evan Hirsch, Bobby Samuel, Jim Atkins, Joseph Golden, Rob Currie, Trevor Hallstein, Kathy Rowerdink, Robert Baertsch, Ted Goldstein, Kristin Zubritzky, Chris Jensen, Beau Norgeot, Chinmay Belthangady, Jennifer Degan, Jason Schlachter, Lee Sanders, and Martin Kline. Thanks, too, to Aaron Shephard; Lauren Wilcox; Paul Bryan and the attendees of UX Strat; Elena Agapie, Daniel Epstein, Madhu Reddy, and the attendees of the UCI Informatics Seminar Series.

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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[2] American Diabetes Association. 2013. Economic Costs of Diabetes in the U.S. in 2012. *Diabetes Care* 36, 4 (03 2013), 1033–1046. https://doi.org/10.2337/dc12-2625

[3] American Heart Association. 2023. Guidelines and Statements. https://professional.heart.org/en/guidelines-and-statements.

[4] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[5] Christopher M. Bell, Arash Jalali, and Edward Mensah. 2013. A Decision Support Tool for Using an ICD-10 Anatomographer to Address Admission Coding Inaccuracies: A Commentary. *Online Journal of Public Health Informatics* 5, 2 (June 2013). https://doi.org/10.5210/ojphi.v5i2.4813

[6] Chinmay Belthangady, Stefanos Giampanis, Ivana Jankovic, Will Stedden, Paula Alves, Stephanie Chong, Charlotte Knott, and Beau Norgeot. 2022. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. *Nature Communications* 13, 1 (Nov. 2022). https://doi.org/10.1038/s41467-022-33732-9

[7] Chinmay Belthangady, Will Stedden, and Beau Norgeot. 2021. Minimizing bias in massive multi-arm observational studies with BCAUS: balancing covariates automatically using supervision. *BMC Medical Research Methodology* 21, 1 (Sept. 2021). https://doi.org/10.1186/s12874-021-01383-x

[8] Andrew B.L. Berry, Catherine Lim, Andrea L. Hartzler, Tad Hirsch, Evette Ludman, Edward H. Wagner, and James D. Ralston. 2017. Creating Conditions for Patients' Values to Emerge in Clinical Conversations: Perspectives of Health Care Team Members. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 1165–1174. https://doi.org/10.1145/3064663.3064669

[9] Andrew B.L. Berry, Catherine Lim, Tad Hirsch, Andrea L. Hartzler, Edward H. Wagner, Evette Ludman, and James D. Ralston. 2017. Getting Traction When Overwhelmed: Implications for Supporting Patient-Provider Communication. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/3022198.3026328

[10] Cynthia M. Boyd, Daniela Vollenweider, and Milo A. Puhan. 2012. Informing Evidence-Based Decision-Making for Patients with Comorbidity: Availability of Necessary Information in Clinical Trials for Chronic Diseases. *PLoS ONE* 7, 8 (Aug. 2012), e41601. https://doi.org/10.1371/journal.pone.0041601

[11] Eleanor R. Burgess, Elizabeth Kaziunas, and Maia Jacobs. 2022. Care Frictions: A Critical Reframing of Patient Noncompliance in Health Technology Design. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 281 (nov 2022), 31 pages. https://doi.org/10.1145/3555172

[12] Eleanor R. Burgess, Madhu C. Reddy, Andrew Davenport, Paul Laboi, and Ann Blandford. 2019. "Tricky to Get Your Head around": Information Work of People Managing Chronic Kidney Disease in the UK. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3290605.3300895

[13] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-Functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 43, 7 pages. https://doi.org/10.1145/3411763.3443435

[15] CDC. 2022. Manage Blood Sugar. https://www.cdc.gov/diabetes/managing/manage-blood-sugar.html.

[16] CDC. 2023. Type 2 Diabetes. https://www.cdc.gov/diabetes/basics/type2.html.

[17] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners.* Sage publications ltd, Los Angeles.

[18] CMS.gov. 2023. Centers for Medicare and Medicaid Services Measures Inventory Tool. https://cmit.cms.gov/cmit.

[19] American Diabetes Association Professional Practice Committee. 2021. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2022. *Diabetes Care* 45 (12 2021), S125–S143. https://doi.org/10.2337/dc22-S009

[20] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* (March 2021), medethics–2020–106820. https://doi.org/10.1136/medethics-2020-106820

[21] Steven Epstein. 1996. *Impure science: AIDS, activism, and the politics of knowledge.* Vol. 7. Univ of California Press.

[22] R. S. Evans. 2016. Electronic Health Records: Then, Now, and in the Future. *Yearbook of Medical Informatics* 25, S 01 (Aug. 2016), S48–S61. https://doi.org/10.15265/iys-2016-s006

[23] U.S. Food and Drug Administration. 2023. Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[24] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, David J. Albers, Arlene M. Smaldone, and Lena Mamykina. 2021. From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 206, 17 pages. https://doi.org/10.1145/3411764.3445555

[25] John H Galla. 2000. Clinical practice guideline on shared decision-making in the appropriate initiation of and withdrawal from dialysis. *Journal of the American Society of Nephrology* 11, 7 (2000), 1340–1342.

[26] Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. 2005. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes. *JAMA* 293, 10 (March 2005), 1223. https://doi.org/10.1001/jama.293.10.1223

[27] Shefali Haldar, Yoojung Kim, Sonali R. Mishra, Andrea L. Hartzler, Ari H. Pollack, and Wanda Pratt. 2020. The Patient Advice System: A Technology Probe Study to Enable Peer Support in the Hospital. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 112 (oct 2020), 23 pages. https://doi.org/10.1145/3415183

[28] George Hripcsak, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, Martijn J. Schuemie, Frank J. DeFalco, Adler Perotte, Juan M. Banda, Christian G. Reich, Lisa M. Schilling, Michael E. Matheny, Daniella Meeker, Nicole Pratt, and David Madigan. 2016. Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences* 113, 27 (June 2016), 7329–7336. https://doi.org/10.1073/pnas.1510502113

[29] Ming-Hui Huang and Roland T. Rust. 2018. Artificial Intelligence in Service. *Journal of Service Research* 21, 2 (Feb. 2018), 155–172. https://doi.org/10.1177/1094670517752459

[30] interactions. 2018. Special Topic: Designing AI. https://interactions.acm.org/archive/toc/november-december-2018/.

[31] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[32] 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 (04 Feb 2021), 108. https://doi.org/10.1038/s41398-021-01224-x

[33] Jonathan H. Jankovic, Ivana; Chen. 2020. Clinical Decision Support and Implications for the Clinician Burnout Crisis. *Yearb Med Inform* 29, 01 (21 Aug 2020), 145–154. https://doi.org/10.1055/s-0040-1701986

[34] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[35] Tessa Kennedy-Martin, Sarah Curtis, Douglas Faries, Susan Robinson, and Joseph Johnston. 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16, 1 (03 Nov 2015), 495. https://doi.org/10.1186/s13063-015-1023-4

[36] Moien Abdul Basith Khan, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi. 2019. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *Journal of Epidemiology and Global Health* 10 (2019), 107–111. Issue 1. https://doi.org/10.2991/jegh.k.191028.001

[37] Vanessa Kizzier-Carnahan, Kathryn A. Artis, Vishnu Mohan, and Jeffrey A. Gold. 2019. Frequency of Passive EHR Alerts in the ICU: Another Form of Alert Fatigue? *Journal of Patient Safety* 15, 3 (2019). https://journals.lww.com/journalpatientsafety/Fulltext/2019/09000/Frequency_of_Passive_EHR_Alerts_in_the_ICU_.12.aspx

[38] Mike Kuniavsky, Elizabeth Churchill, and Molly Wright Steenson. 2017. Designing the User Experience of Machine Learning Systems. https://mikekparc.github.io/AAAI-UX-ML/.

[39] Catherine Lim, Andrew B.L. Berry, Tad Hirsch, Andrea L. Hartzler, Edward H. Wagner, Evette Ludman, and James D. Ralston. 2016. "It Just Seems Outside My Health": How Patients with Chronic Conditions Perceive Communication Boundaries with Providers. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 1172–1184. https://doi.org/10.1145/2901790.2901866

[40] Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F. De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J. Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, and Andrew Georgiou. 2019. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 28, 01 (16 Aug 2019), 128–134. https://doi.org/10.1055/s-0039-1677903

[41] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: Investigation of Social Scaffolding for Reflective Thinking in Diabetes Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/1357054.1357131

[42] Mohammad. Mazaheri Habibi, Fahimeh. Abadi, Hamed. Tabesh, Hasan. Vakili-Arki, Ameen. Abu-Hanna, and Saeid. Eslami. 2018. Evaluation of patient satisfaction of the status of appointment scheduling systems in outpatient clinics: Identifying patients' needs. *Journal of Advanced Pharmaceutical Technology & Research* 9, 2 (2018), 51–55. https://doi.org/10.4103/japtr.JAPTR_134_18

[43] Victor M. Montori, Amiram Gafni, and Cathy Charles. 2006. A shared treatment decision-making approach between patients with chronic conditions and their clinicians: the case of diabetes. *Health Expectations* 9, 1 (2006), 25–36. https://doi.org/10.1111/j.1369-7625.2006.00359.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1369-7625.2006.00359.x

[44] Karen C Nanji, Diane L Seger, Sarah P Slight, Mary G Amato, Patrick E Beeler, Qoua L Her, Olivia Dalleur, Tewodros Eguale, Adrian Wong, Elizabeth R Silvers, Michael Swerdloff, Salman T Hussain, Nivethietha Maniam, Julie M Fiskio, Patricia C Dykes, and David W Bates. 2017. Medication-related clinical decision support alert overrides in inpatients. *Journal of the American Medical Informatics Association* 25, 5 (10 2017), 476–481. https://doi.org/10.1093/jamia/ocx115 arXiv:https://academic.oup.com/jamia/article-pdf/25/5/476/34149920/ocx115.pdf

[45] National Institute of Diabetes, Digestive, and Kidney Diseases. 2017. Type 2 Diabetes. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes.

[46] Nazmun Nisat Ontika, Hussain Abid Syed, Sheree May Saßmannshausen, Richard HR Harper, Yunan Chen, Sun Young Park, Miria Grisot, Astrid Chow, Nils Blaumer, Aparecido Fabiano Pinatti de Carvalho, et al. 2022. Exploring Human-Centered AI in Healthcare: Diagnosis, Explainability, and Trust. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

[47] Jerome A. Osheroff, Jonathan M. Teich, Blackford Middleton, Elaine B. Steen, Adam Wright, and Don E. Detmer. 2007. A Roadmap for National Action on Clinical Decision Support. *Journal of the American Medical Informatics Association* 14, 2 (03 2007), 141–145. https://doi.org/10.1197/jamia.M2334 arXiv:https://academic.oup.com/jamia/article-pdf/14/2/141/2563507/14-2-141.pdf

[48] Jonathan M. Osheroff, Jerome A. and Teich, Luis Levick, Donald and Saldana, Ferdinand T. Velasco, Dean F. Sittig, Kendall M. Rogers, and Robert A. Jenders. 2012. *Improving Outcomes with Clinical Decision Support: An Implementer's Guide, Second Edition*. Taylor and Francis.

[49] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. https://doi.org/10.1145/3411763.3441347

[50] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 51, 22 pages. https://doi.org/10.1145/3491102.3517672

[51] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S. Lasecki. 2019. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 506–510. https://doi.org/10.1145/3311957.3359433

[52] Ian Paterson, Krishnan Ramanathan, Rakesh Aurora, David Bewick, Chi-Ming Chow, Brian Clarke, Simone Cowan, Anique Ducharme, Kenneth Gin, Michelle Graham, Anil Gupta, Davinder S. Jassal, Mustapha Kazmi, Andrew Krahn, Yoan Lamarche, Ariane Marelli, Idan Roifman, Marc Ruel, Gurmeet Singh, Larry Sterns, Ricky Turgeon, Sean Virani, Kenny K. Wong, and Shelley Zieroth. 2021. Long COVID-19: A Primer for Cardiovascular Health Professionals, on Behalf of the CCS Rapid Response Team. *Canadian Journal of Cardiology* 37, 8 (01 Aug 2021), 1260–1262. https://doi.org/10.1016/j.cjca.2021.05.011

[53] Justin Porter, Cynthia Boyd, M. Reza Skandari, and Neda Laiteerapong. 2023. Revisiting the Time Needed to Provide Adult Primary Care. *Journal of General Internal Medicine* 38, 1 (01 Jan 2023), 147–155. https://doi.org/10.1007/s11606-022-07707-x

[54] Professional Practice Committee. 2020. Professional Practice Committee: Standards of Medical Care in Diabetes—2021. *Diabetes Care* 44 (12 2020), S3–S3. https://doi.org/10.2337/dc21-Sppc arXiv:https://diabetesjournals.org/care/article-pdf/44/Supplement_1/S3/551668/dc21sppc.pdf

[55] Nancy Robert. 2019. How artificial intelligence is changing nursing. *Nursing Management* 50, 9 (2019). https://journals.lww.com/nursingmanagement/Fulltext/2019/09000/How_artificial_intelligence_is_changing_nursing.8.aspx

[56] Tracy L. Setji, Courtne y. Page, Neha Pagidipati, and Benjamin A. Goldstein. 2019. Differences in Achieving Hba1C Goals Among Patients Seen by Endocrinologists and Primary Care Providers. *Endocrine Practice* 25, 5 (01 May 2019), 461–469. https://doi.org/10.4158/EP-2018-0405

[57] Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C.Cordell Green, and Stanley N. Cohen. 1975. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* 8, 4 (1975), 303–320. https://doi.org/10.1016/0010-4809(75)90009-9

[58] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 32, 11 (2021), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

[59] Matt Turek. 2023. Explainable Artificial Intelligence (XAI). https://www.darpa.mil/program/explainable-artificial-intelligence.

[60] Tiffany C. Veinot, Kai Zheng, Julie C. Lowery, Maria Souden, and Rosalind Keith. 2010. Using Electronic Health Record Systems in Diabetes Care: Emerging Practices. In *Proceedings of the 1st ACM International Health Informatics Symposium* (Arlington, Virginia, USA) *(IHI '10)*. Association for Computing Machinery, New York, NY, USA, 240–249. https://doi.org/10.1145/1882992.1883026

[61] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages.

https://doi.org/10.1145/3476068

[62] Darina Vorobeva, Yasmina El Fassi, Diego Costa Pinto, Diogo Hildebrand, Márcia M. Herter, and Anna S. Mattila. 2022. Thinking Skills Don't Protect Service Workers from Replacement by Artificial Intelligence. *Journal of Service Research* 25, 4 (2022), 601–613. https://doi.org/10.1177/10946705221104312 arXiv:https://doi.org/10.1177/10946705221104312

[63] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 697, 18 pages. https://doi.org/10.1145/3411764.3445432

[64] Lauren Wilcox, Rupa Patel, Anthony Back, Mary Czerwinski, Paul Gorman, Eric Horvitz, and Wanda Pratt. 2013. Patient-Clinician Communication: The Roadmap for HCI. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 3291–3294. https://doi.org/10.1145/2468356.2479669