Patient Record Summarization Through Joint Phenotype Learning and Interactive Visualization

Gal Levy-Fix

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Patient Record Summarization Through Joint Phenotype Learning and Interactive Visualization

Gal Levy-Fix

Complex patient are becoming more and more of a challenge to the health care system given the amount of care they require and the amount of documentation needed to keep track of their state of health and treatment. Record keeping using the EHR makes this easier but mounting amounts of patient data also means that clinicians are faced with information overload. Information overload has been shown to have deleterious effects on care, with increased safety concerns due to missed information. Patient record summarization has been a promising mitigator for information overload. Subsequently, a lot of research has been dedicated to record summarization since the introduction of EHRs. In this dissertation we examine whether unsupervised inference methods can derive patient problem-oriented summaries, that are robust to different patients. By grounding our experiments with HIV patients we leverage the data of a group of patients that are similar in that they share one common disease (HIV) but also exhibit complex histories of diverse comorbidities. Using a user-centered, iterative design process, we design an interactive, longitudinal patient record summarization tool, that leverages automated inferences about the patient's problems. We find that unsupervised, joint learning of problems using correlated topic models, adapted to handle the multiple data types (structured and unstructured) of the EHR, is successful in identifying the salient problems of complex patients. Utilizing interactive visualization that exposes inference results to users enables them to make sense of a patient's

problems over time and to answer questions about a patient more accurately and faster than using the EHR alone.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Noémie, thank you for these last five years. From our first meeting I had a sense we would get along swimmingly. Thank you for pushing me when I needed to be pushed and for giving me space to do my thing. Your guidance and passion was exactly what I needed to get this done. To my committee members, George, Gil, David, and Mike, I'm honored and thankful for you serving on my dissertation committee.

Thank you to my family. Eran, your support and devotion made all this possible. I so love achieving things with you by my side. Can't wait for this next adventure together. To my patents, thank you for your unconditional love, constant encouragement, and oh so needed childcare in times of need. To my little ones, Bay and Sky, thank you for driving me nuts and yet keeping me sane at the same time.

Thank you to my classmates and friends. Alex and Amelia, thank you for skyping me in to classes while I was on maternity leave. Thank you to my lab mates and DBMI fellow students for being awesome, caring, funny, nerdy people. Thank you to Marina, Sharon, and Emma for always being willing to help. Our random chats in the hallways of DBMI made the trip in worth it. Iñigo and Victor, thank you for the (almost) weekly lunches. Your friendship and laughs made me feel almost normal during crazy times. Thank you to Jason, my clinical collaborator and friend. Without you I would not be graduating on time.

Dedication

To my boys, Eran, Bay, and Sky, thank you for being my constant reminder of what matters.

Chapter 1: Introduction

1.1 The need for longitudinal patient summaries

Most adults in the US today live with at least one chronic disease [1]. Many of those suffer from multimorbidty, the existence of more than one disease [2]. With an aging population the number of complex patients is on the rise [3]. Chronic disease and multimorbidty puts a high burden on patients and also the healthcare system, with it being a leading driver of health care costs in the US [4]. These patients often have long medical histories, with high utilization of health services, that is complicated by simultaneous and interconnected disease processes. With the ubiquitous adoption of electronic medical records, patient data is extensively recorded. However, with overwhelming amounts of historical patient data, clinicians experience significant burden when needing to sift through the patient record to get understanding of the patient case [5]. This is especially true for patients with chronic problems for which temporality plays a large role [6]. Current electronic health record (EHR) systems remain very visit oriented, with limited support for temporal views (especially beyond one data source).

To support clinicians in making use of historical data there is a need for clinical decision support (CDS) that help clinicians effectively review and digest patient data, transforming data into actionable insight. Most adopted CDS systems focus on generating alerts and recommendations but few help clinicians make sense of patient data at the bedside. The task of establishing an accurate and comprehensive mental image of the patient is particularly daunting for complex patients. Difficulties include understanding the chronology of problems, symptoms, and treatment as well as identifying how problems relate to one another [6].

Existing approaches to patient record summarization have focused on the problem oriented record [7–9]. However few propose tools that facilitate clinicians ability to identify problems, how

they relate, and change overtime for a given patient. In order to effectively support clinicians in these tasks there is a need for solutions that leverage methods in machine learning in order for summarization to be robust and generalizable to diverse types of problems and patient complexities. Moreover, summarization solutions can benefit greatly from the utilization of visualizations to facilitate quick identification of patterns and interactivity to allow for drilling down to patient details to facilitate trust. Finally summarization approaches that leverages machine learning and visualization in combination needs to be validated as a whole in realistic clinical settings to prove clinical usefulness for realistic tasks to encourage adoption.

1.2 Selected use case - Human Immunodeficiency Virus (HIV) patients

An extreme example of patients that suffer from chronic disease and multimorbidty are patients with HIV. As a working example for complex patients, we elected to focus the work in this dissertation on record summarization of patient with HIV. HIV is a chronic disease which requires constant care and surveillance and is also associated with a high number of comorbidities [10]. Similar to other patients with chronic disease and multimorbidty patients have long and complex medical histories which are difficult to track and are important to consider for effective treatment decisions.

High multimorbidty rates in this population can be attributed to several factors. With the success of Antiviral treatment (ART) HIV patient commonly live into their 70s and thus suffer from multimorbidty that are associated with aging. The development of multimorbidty associated with aging are caused through the low-grade inflation that is developed in older age. The same mechanism of disease is reinforced through the chronic immune activation that occurs in HIV is suspected to drive to expedited development of age-related multimorbidty [11]. Finally, ART toxicity is also associated with increased risk of certain multimorbidty. HIV patients have been found to be at higher risk for cardiovascular disease, renal disease, osteoporosis, metabolic disorders, and several cancers [10]. Thus treatment decision and effective treatment of HIV patients focuses much beyond viral-suppression and requires a holistic understanding of patient multimorbidty over time

[10]. Moreover, due to high healthcare utilization, patients have complex and long medical histories for which patient data review is particularly burdensome and exacerbates the need for summarization. The intention is that the findings regarding longitudinal summarization of patients with HIV will be able to generalize to other patient types that also exhibit multimorbidty and require constant, chronic care.

1.3 Thesis approach

The studies in this thesis investigate the use of unsupervised computational methods and interactive visualization to support automated longitudinal summarization of patient records with chronic multimorbidty. Design and computational requirements for the automated summarization are collected through an iterative user-centered design approach. The thesis describes the information collection of target users, and several cycles of design requirements refinements, prototype development, and evaluation of the proposed patient summarizer. As a generalizable use case, the experimental design focuses on supporting clinicians from the HIV clinic at NewYork-Presbyterian Hospital (NYPH) as potential users of the purposed summarizer system. Patient data of HIV positive patients are used to train and evaluate the computational model used for summarization. Aim 1 of this thesis describes the iterative user-centered-design approach used to come up with design requirements for longitudinal patient summarization system; Aim 2 includes the computational model development and evaluation for patient record summarization; and Aim 3 investigates the use of the proposed summarization to support the clinical task of patient chart review through a task-based experimental usability study with target users. All studies presented in this dissertation using patient data or clinician participants were approved by the Columbia University Institutional Review Board.

1.3.1 AIM I: Collecting design requirements through iterative user-centered design

Objective: Collect design requirements for longitudinal patient record summarization that supports the information needs of HIV clinicians when reviewing patient data for the purpose of

clinical care.

Hypotheses:

- $_{0\text{-}1}$: A problem oriented view of patient record augmented with how problems relate and change over time is useful for clinicians in the context of HIV patient care, specifically in patient chart review for new and existing patients
- $_{0\text{-}2}$: An interactive sankey diagram visualization of patient problems, how they change, and relate overtime is an appropriate way to showcase patient information
- $_{0\text{-}3}$: The proposed visualization helps clinicians construct patient problem lists

Methods and materials: Aim 1 of this thesis relies on an iterative user-centered design approach to come up with design requirements for longitudinal patient summarization system (Figure 1.1). Information needs for longitudinal patient chart review are collected from published literature. Semi-structured interviews with HIV clinicians are utilized to confirm that the identified information needs from the literature are consistent with the needs of the selected use case for this thesis. An initial prototype of the system is developed and tested in a preliminary user study. Information needs collected from the literature, the clinician interviews, and the preliminary user study are translated to design requirements of the patient summarizer system.

Primary findings: The information needs collected from the literature identified that in the chart review of patient longitudinal data clinicians seem to i) identify the patient main problems, ii) verify those problems with alternative data sources in the patient record; and iii) ascertain the status of the patient problem. Clinicians that care for patients with chronic disease and multimorbidity, look to identify inter-relatedness in the data to support the complex state of their patients. The information needs of clinicians that care for HIV patients were found to be consistent with those identified in the literature. Interviews with HIV clinicians emphasized the importance of understanding the patient's state of comorbidities that dominate a lot of their care plans. The preliminary

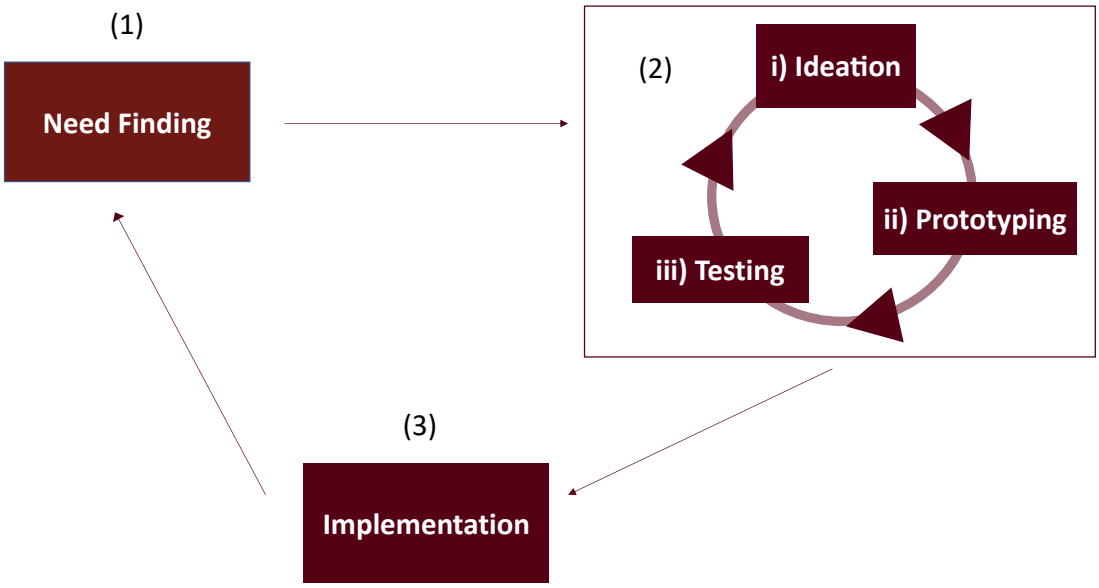

Figure 1.1: Diagram of design process. Figure adapted from [12].

user study indicated that showcasing patient main problems using sankey diagram is a reasonable representation of patient problems and their evolving dominance over time. Information needs collected were translated into 6 design requirements for the summarization. Design requirements are operationalized through a combination of interface design, computational method requirements, and system interactivity.

1.3.2 AIM II: Summary and content selection through joint phenotype learning

Objective: Develop and evaluate computational model that infers patient problems and problem correlations through time while preserving data provenance.

Hypotheses:

- $_{0\text{-}1}$:Proposed model is able to learn clinically interpretable phenotypes
- $_{0\text{-}2}$:Proposed model is able to learn HIV and non-HIV phenotypes using the EHR data of a patient cohort with many comorbidities
- $_{0\text{-}3}$:Proposed model is able to learn *clinically valid* relationships between phenotypes

- $_{0\text{-}4}$:Proposed model is able to learn *diverse types* of relationships between phenotypes

Methods and materials: In Aim 2 a joint computational phenotyping approach is developed and utilized to identify patient problems and correlations between problems by training on structured and unstructured data from the electronic health records of a large patient cohort of HIV patients. Previously proposed variational inference approach is extended to support multi-source data. Learned phenotypes and phenotype-relationships identified from the patient population are qualitatively evaluated for clinical validity by two clinical experts. Coverage and diversity of the learned phenotypes and relationships are quantitatively compared to an available knowledge-based baseline. The phenotyping approach allows for inference on a single patient record to generate a patient-level summary of problems over time.

Primary findings: According to a qualitative evaluation by two clinicians of the proposed model output, the model is able to identify clinically meaningful phenotypes and phenotype-relationships. The method identifies several HIV sub-phenotypes as well as many non-HIV phenotypes. The non-HIV phenotypes are found to be diverse and cover a wide range of comorbidities as seen when comparing to a knowledge-base baseline that classifies diagnosis codes in to disease groups. Phenotype relationships found are largely consistent with the knowledge baseline, with several relationships identified to be clinically correct but that could not be inferred by the baseline. When the model is applied on a single patient record it provides a summary of the patient phenotypes and the data that was assigned to each phenotype. Patient data is found to be well represented by 1-10 phenotypes, providing significant dimensionality reduction to patient data that could be digested by clinicians at the point of care.

1.3.3 AIM III: Usability testing of patient longitudinal summarizer

Objective: Implement and evaluate clinical utility of patient summarizer tool which on the back-end leverages probabilistic model to automatically learn patient problems and on the front-end displays problems in a web interactive environment.

Hypotheses:

- $_{0\text{-}1}$:Implemented tool is found to aid clinician in obtaining a mental image of the patient
- $_{0\text{-}2}$:Principles from human-centered machine learning assist clinicians improves the usability of a machine learning CDS system

Methods and materials: Aim 3 integrates the identified design requirements from Aim 1 and the computational model from Aim 2 into a single system that summarizes and presents new patient data in a secure interactive web environment. The summarizer utility in support patient chart review is evaluated in a mixed-methods user study with clinicians simulating real world conditions. The user study compares the performance of subjects in chart review tasks using the summarizer to the EHR baseline. Under each study condition participants are asked to construct the patient problem list, answer two clinical questions regarding the patient, and generated a short summary of the patient case. Participants are asked to fill a standard usability survey and provide their free form feedback regarding their experience with the tool.

Primary findings: Clinicians spent slightly more time constructing patient problem lists with the summarizer, but generated lists with higher recall. The summary was found to increase question-answer accuracy and reduce the time-to-completion significantly. A usability survey found that half the users stated that they would use the system frequently. Most of the users found the system consistent and easy to use. The visual representation of the patient problem over time using sankey diagrams was well received and users found to be an intuitive snapshot of patient state over time. Change in problem line thickness was found helpful in identifying new problems or flare up of certain diseases. Users however did not want to relay on line thickness for certain diseases for which laboratory results are clear indication of the disease state, such as HIV. Feedback from the participants identified that some clinicians thought the summary provided enough information to help them judge the accuracy of the automated summary, while others showed very low tolerance for inconsistencies in the system. Several users indicated they wished the summary was directly linked to more components in the patient EHR such as notes, laboratory test results, procedures,

and medication dosages.

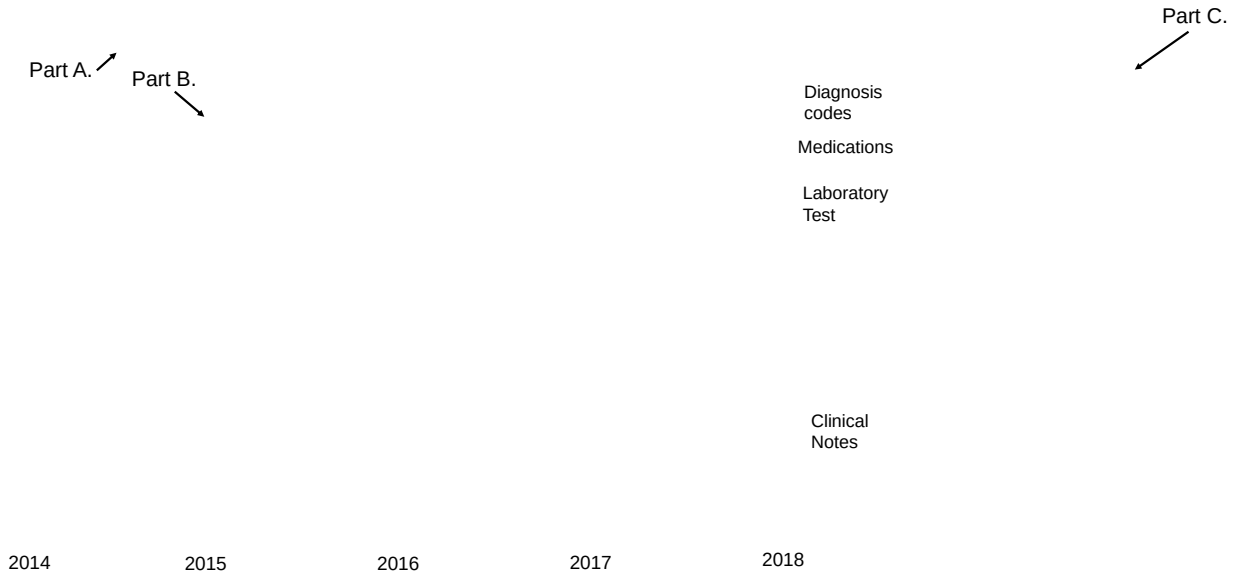

Figure 1.2: Example of patient-specific summary over five years. The top 7 most salient problems in between 2014 and 2018 are visualized and how their documentation has evolved through time. The summary is presented at the year level by binning the patient's documentation for that time resolution. The patient has HIV-specific problems, as well as comorbidities, including asthma, depression, and substance abuse. Relations among the inferred phenotypes are not shown. Dates are changed to maintain patient privacy.

1.4 Contributions

This dissertation contributes a literature review regarding the use of machine learning and data visualization for CDS and identifies gaps and opportunities for these methods in three main types of CDS. We contribute the design requirements of an information tool that leverages visualization, machine learning, and interactivity. We further identify that sankey diagrams are a viable visualization for patient problems and their evolution over time. The dissertation further provides a methodological contribution for joint learning of phenotypes and phenotype relationships using EHR data. We extended previously proposed technique for variational inference for non-conjugate distributions to allow for multiple inputs in the model. The proposed joint learning of phenotypes and their relationships can be used for characterization of patient populations, for data driven phenotype relationship discovery, for patient summarization, and for down stream tasks such as patient

level prediction. Finally, the dissertation demonstrates that coupling of interpretable unsupervised machine learning and interactive visualization has the promise to support clinicians in patient chart review.

1.5 Guide for the Reader

Chapter 2 outlines previous research on various types of CDS and the methodologies utilized for those systems. The chapter outlines gaps and opportunities of using ML and data visualization for patient summarization but also other types of CDS. The chapter also describes recent work on Human Centered Machine Learning.

Chapter 3 describes our process to identify the informational needs of clinicians when reviewing patient longitudinal data. We describe our findings and conduct interviews with HIV clinicians to confirm their needs align with findings from the literature. We further describe a preliminary study, evaluating the appropriateness of using sankey diagrams to represent patient problems over time in a task based evaluations. We conclude the chapter by translating the identified information needs to design requirements of a summarization system.

Chapter 4 we describe our work on the development and evaluation of a computational model and it's inference that leverage unsupervised probabilistic machine learning to perform interpretable dimensionality reduction of patient data, both structured and unstructured.

Chapter 5 reports on the construction of a summarization system that leveraged the design requirements identified in Chapter 3 and the model described in Chapter 4. The chapter describes the experimental design of an evaluation study with clinicians to assess the usability of the proposed summarization system to support them in patient chart review of HIV patients.

Chapter 2: Background

2.1 Clinical decision support [1]

The grand vision of a learning health system hold the promise for providing more personalized, higher quality, safer, and efficient care [13]. The learning health system pipeline involves systematically gathering clinical data, learning from that data and generating evidence, and feeding it back to clinicians in real-time to help with decision making. This process highly depends on the robust development, evaluation, and adoption of various types of CDS in clinical care to deliver knowledge to the point of care. However, for CDS to help realize the goals of a learning health system, numerous challenges have to be addressed. Challenges to the effective use of CDS include not being sufficiently patient-specific, utilizing simplistic CDS logic, lacking generalizability, and failing to address human factor issues [14].

There is a growing interest in medicine to leverage machine learning for clinical decision support [15]. However, there has have been limited examples where machine learning based systems have been used in the clinic. Bottlenecks of implementing machine learning based approaches in the clinic have to do with accuracy and validation but many aspects also relate to the usability of these systems, tying to user-centered design [16]. To overcome these challenges user-centered design principles could be leveraged to make more user focused machine learning systems. Research at the intersection of people's needs and machine learning has had a growing interest in recent years and has been referred to as human-centered machine learning (HCML). The premise of this body of work is refocusing machine learning from a human goals perspective [17]. Outside of the healthcare domain, emerging research in HCML includes several diverse sub fields such as human-in-the loop approaches [18–20], method interpretability and explainability [21–23], and

[1] A large part of this chapter will appear in the Annual Review of Biomedical Data Science, Vol 4, 2021.

fairness [24]. Healthcare oriented research on the matter has focused on calling for smart systems that are better aligned with clinical task [25], method interpretability [26], and fairness [27].

Leveraging recent developments in machine learning and data visualization, especially in combination under this paradigm of HCML, could help overcome previously cited challenges of CDS and machine learning based CDS in particular. Machine learning methods have the potential to enhance CDS tools by generating new knowledge from gathered data, providing better patient specificity, supporting the identification of complex patterns, and improving generalizability to different patients and conditions. Data and information visualization (dataVis) techniques, from static to interactive visualizations to more complex visual analytics, have the potential to assist with feeding back information to clinicians and improve the interpretability and transparency of CDS systems. In this way machine learning and data visualization provide complimentary benefits to CDS and may be synergistic in combination (Figure 2.1). Thus, there is a strong case for greater focus on leveraging machine learning and data visualization in combination to help the realization of a learning health system.

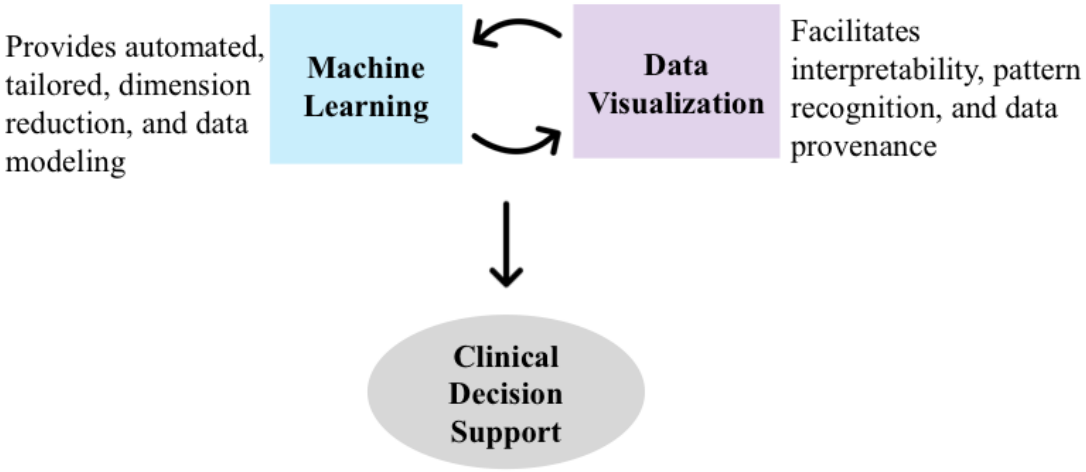

Figure 2.1: Synergy of data visualization, machine learning, and clinical decision support. This chapter is dedicated to describing and synthesizing the current state of the literature on machine learning and data visualization methods used for clinical decision support.

This review is based on a survey of the CDS literature and literature describing methodology developed for CDS applications. Pool of papers was compiled through a search on PubMed for

the terms “clinical decision support”, “machine learning”, or “visualization.” In addition, papers were compiled from machine learning for health conference proceedings (Machine Learning for Healthcare, NeurIPS Machine Learning 4 Health Workshops) and IEEE Visualization conference proceedings. We focused our search to publications from 2010 to 2019, but also included earlier seminal works dating back to 1959. Additional papers were identified through “pearl growing,” until we reached thematic saturation. We restricted our focus to papers describing clinician-facing clinical decision support, whether patient-specific or cohort-level, and utilizing EHR data collected through clinical documentation. Papers were classified into three general types of CDS, although gaps and opportunities from analysis of current work in these CDS types should also generalize to other CDS types. The three CDS types described in this chapter are Infobutton, Content Summarization and Organization, and Alert [28].

1. **Infobuttons** are a type of CDS developed to help clinicians retrieve external resources relevant to the care of their patients such as scientific publications and guidelines. As medical evidence is constantly generated and updated and as clinicians have less time at the point of care, Infobuttons make it easier to stay up to date and well informed.

2. **Content Summarization and Organization (CSO) CDS** is used to summarize or re-organize patient-level or cohort-level information to clinicians in a way that facilitates understanding, pattern recognition, and decision making. Current EHR systems contain large amounts of data even for single patients, making the tasks of information gathering and synthesis cognitively difficult and time consuming. CSO CDS aims to help centralize and crystallize patient data available for better and easier decision making.

3. **Alert CDS** provides alerts, reminders, and recommendations in the context of patient data, clinician actions (such as medication orders), and clinical knowledge. Due to the high volume of data available in the EHR, limited clinician time, and evolving medical guidelines, clinicians may miss important information regarding a given patient that could lead to better and safer care. Alert CDS produces a single output such as a prediction, an alert, or a set of

recommendations to clinicians in order to help direct action and prevent medical errors.

This chapter synthesizes previous work on each CDS type (Infobuttons, CSO CDS, and Alert CDS) and the machine learning and data visualization methods utilized. Literature on each CDS type is grouped and described by the type of methods they utilize (Figure 2.2). We review how each CDS type has applied: (1) heuristics-based knowledge development methods (heuristics) defined to be expert curated rules or knowledge-based sources such as ontologies; (2) machine learning (ML) defined to be data-driven and learning-based methods for knowledge development; and (3) data visualization (dataVis) defined to be the advanced visual representation of data and information using static or interactive graphs, diagrams, or pictures to convey information; and (4) any combination of these three methods. This thesis proposes to combine the use of machine learning and visualization to construct a patient record summarization tool, a type of CSO CDS. This section synthesizes why, when, and how machine learning and visualization are used in previous works in each CDS type. Works outside of the CDS domain that use machine learning and visualization that can inform our work are also reviewed here.

2.1.1 Infobutton clinical decision support

Infobuttons are systems developed to help clinicians retrieve external resources such as scientific publications and guidelines that are relevant to their patients and informational needs. As medical evidence is constantly being generated and updated and with clinicians having less available time, Infobuttons make the task of accessing up to date medical evidence relevant to their clinical cases easier.

Types of methods for Infobutton CDS

Of the three CDS types, Infobuttons represent the most common CDS implemented in the EHR [28]. Research on heuristics-based Infobuttons, with most work taking place in the 1990s, leverages a combination of ontological knowledge and rules to identify clinical concepts in patient

Figure 2.2: Venn diagram showcasing the intersections between clinical decision support, machine learning, and data visualization. We refer to heuristics-based methods as rules that are expert-curated or that rely on knowledge sources such as ontologies. Machine-learning methods include clinical data-driven methods. Visualization methods include static, interactive, as well as advanced visual analytics from clinical data.

records and construct relevant search queries to look for relevant resources in scienti c article databases or online [29–36].

Infobuttons leveraging machine learning largely focus on the personalization and summarization of the outside resources retrieved by the system and returned to clinicians. These systems largely represent experimental stand-alone systems that have not been integrated in EHR systems, as have traditional Infobuttons. General approaches of these works include context aware scienti c article summarization, recommendation of outside resources based on patient data, learning to rank models of articles based on clinician search queries, and question-answering related to a patient's clinical case [37–41].

Very limited works have leveraged data visualization for Infobutton CDS in isolation or in combination with machine learning. The limited work in this area has looked at an interactive citation screening system for improved clinical question answering [42].

Table 2.1: Examples of Infobutton CDS by method type

Papers	Type of method utilized	Description
Powsner et al. 1989 [30]	Heuristics	Utilize rule-based MEDLINE searches by clinical topic
Cimino et al. 1997 [35]	Heuristics	Use terminology knowledge (Medical Entities Dictionary) to select queries and resources
Elhadad et al. 2005 [37]	ML	Use Natural language processing to tailor summaries of scienti c articles based on the clinical context of the patient. Evaluated via user study of simulated clinical task compared effectiveness of tailored summary, to non-tailored summary and general article search
Monteiro et al. 2015 [38]	ML	Recommender system of reports and studies based on patient information and clinical context
Donoso-Guzmán and Parra 2018 [42]	ML+dataVis	Compare two relevance feedback algorithms, Rocchio and BM25, in an interactive visualization for citation screening. Evaluated ef ciency and effectiveness of tool in citation screening in user group
More papers by method category:	Heuristics	[29, 31, 33, 34, 43]
	ML	[40, 41]

Gaps and opportunities for ML and dataVis in Infobutton CDS

Results of heuristics-based Infobuttons may still return large amounts of content for clinicians to review in order to find relevant information for their patients. For instance, scientific literature about a specific clinical concept might return hundreds of highly relevant publications. Furthermore, most Infobuttons search resources for one piece of information in the patient record and does not consider combination of clinical concepts, potentially reducing the relevance and usefulness of the retrieved sources. One way to reduce the complexity and size of results is to curate content based on clinical expertise, but this might limit the scope of Infobuttons as well as their sustainability as new evidence emerges.

Data-driven methods can be used to organize further the results of Infobuttons, whether to condense and synthesize the evidence or to personalize and tailor the evidence to the clinician's information needs and clinical context, thus making the information search quicker and more efficient for the clinician [37]. An additional promise for data-driven Infobuttons, rather than rule-driven ones is increased generalizability to different types of searches and concepts with less reliance on manual curation of content. Supervised machine learning solutions, however still require annotated data sets which in the clinical context can be time-consuming and expensive to obtain. It is also important to note that data-driven systems have so-far been mostly evaluated for accuracy and effectiveness in a laboratory setting outside of a deployed, real-world setting. More research is required to evaluate their utility and performance in clinical settings. The lack of integration of visualization in this line of research is also a missed opportunity. Work outside of the health domain has shown that data visualization can help users identify relevant information in information retrieval tasks and facilitate thematic analysis of large sets of documents [44–47].

2.1.2 Clinical Summarization and Organization (CSO) clinical decision support

CSO clinical decision support systems either summarize or re-organize patient information. Dashboards, that select specific data points and presents them in a centralized way, or summarizers which synthesize entire records belong to this type of CDS. These systems aim to help users with

information that is dif cult to digest in its original form in the EHR due to its volume, complexity, or it scattered nature in the EHR.

Type of methods for CSO decision support systems

Literature on CSO decision support largely leverage heuristics-based methods such as expert curated variable selection and knowledge-based sources to organize [48, 49] and summarize patient information [50–58]. These systems have mainly focused on extractive summaries [59] which extract selected information from the patient record into condensed tables,[50–56]. Fewer works have provided abstractive summaries [59] which reformulate patient's data. Those that have generated abstractive summaries of patient data have largely done so by automatically inferring patient problem lists using structured data and supervised machine learning methods [58, 60–62].

Several data visualization technique have also been proposed in combination with heuristics-based CSO systems. Popular approaches for visual extractive summaries have been small visuals and patient data temporal views [63–74]. A few examples also exist of visualizations of abstractive summaries or reorganization of selected patient data [75–85].

Machine learning used for CSO systems have generated abstractive summaries of the patient rather than extractive summaries. That is, reducing patient data dimensionality and complexity into more salient, condensed, and digestible form. These approaches have included automatically generating short narrative description of patients' data and generating the patient's problem list using natural language processing methods (NLP) and supervised machine learning methods [62, 86–92]. Another machine learning approach that has been used to reducing patients' data dimensionality is computational phenotyping. Although most commonly proposed for feature engineering for downstream predictive tasks, interpretable abstraction of patient's clinical data from data driven phenotyping could also be used for patient summarization in clinical decision support [93]. Computational approaches that propose data-driven phenotyping include probabilistic models [93–98], deep learning,[99–104] clustering [105], and decision trees [106].

Few examples in the literature have used a combination of machine learning and visualization

methods for CAO systems. One group of works focus on leveraging machine learning and interactive visualization to showcase cohort visualizations aimed to assist clinicians with patient-level decision making. These works mostly divide into performing two tasks: 1) computation of patient sequence similarity using different clustering methods [107–109]; and 2) frequent patterns identi cation using advance association rules and latent model methods [110–112]. Another group of work leverage the machine learning and visualization for patient-level visualization [113–115]. Some of these works have focus on generating abstractive summaries of patients' data using semi-supervised and unsupervised methods and visualizing those abstractions[114, 115]. Evaluation methods utilized for systems leveraging machine learning and visualization include usability studies with clinical experts [107, 109, 110, 112, 113], interactivity performance [107], and prediction performance using patient-level abstractions [114, 116]

Gaps and opportunities for ML and dataVis in CSO clinical decision support

Heuristics-based CSO systems have been shown to improve physicians' information retrieval capabilities, reduce information overload, improve patient outcomes, and increase guideline compliance[5, 48, 50, 52, 53]. However, such systems often focus on one condition at a time and require extensive expert input and thus are hard to scale to many patient and many disease types. Moreover, these systems mostly focus on extractive summaries which may still contain overwhelming amount of information and thus do effectively alleviate the problem of information overload [127]. Furthermore, a lack visualization use can limit in the effectiveness of the proposed summaries [83].

Machine learning based patient summarization of patient records using NLP have mostly focused on non-temporal summarization [56] or have leveraged only clinical text [78]. These methods are largely extractive and still suffer from many limitations including identifying concept similarities, handling temporality, handling data missingness, identifying importance, leveraging existing knowledge, and deployment [128]. Introduction of machine learning methods that are unsupervised and high-throughput automate dimensionality reduction of complex patient data into

Table 2.2: Examples of CSO Clinical decision support by method type

Papers	Type of method utilized	Description
Alkesic et al. 2017 [49]	Heuristics	Organize clinical content using manual tagging of EHR content for chronic disease tracking
Meystre and Haug 2006 [57]	Heuristics	Infer patient problems using knowledge-based sources
Powsner and Tufte 1994 [63]	dataVis	Patient record summary using small graphs showing laboratory results, medications, vitals, and imaging
Bui et al. 2007 [80]	dataVis	Problem centric patient record temporal abstractive summary using knowledge-based source
Van Vleck and Elhadad 2010 [91]	ML	Natural language processing and classi cation to predict problem relevance for clinical summarization. Automated patient problem summaries compared to expert generated gold standard
Joshi et al. 2016 [116]	ML	Learning identi able patient phenotypes using non-negative matrix factorization. Qualitative evaluation of clinical expert of learned phenotypes and performance in mortality prediction
Guo et al. 2018 [110]	ML+dataVis	Use tensor decomposition to identify latent evolutions of care sequences. Present threads of latent sequences in treatment sequences
Joshi et al. 2012 [115]	ML+dataVis	Utilize novel clustering algorithm to generate layered-grouping of patient states. Real time visual of patient severity by organ system during ICU stay
More papers by method category:	Heuristics	[48, 50, 51, 53–58]
	ML	[87–89, 91, 93–96, 98, 105, 106, 117–121]
	dataVis	[8, 63, 64, 66, 68–82, 84, 85, 122–126]
	ML+dataVis	[107–109, 111, 114]

abstractive summaries that utilize more information from the patient record relative to extractive summaries with little or no human input [93, 99, 129, 130]. However, few works in this area have been investigated speci cally for CSO systems and often do not consider any aspect of HCML, such as usability and interpretability of model output for clinicians.

The use of data visualization have been shown to support pattern identi cation across patient parameters and time [69] While visual summaries of patients' raw data preserves data provenance which can strengthen trust in the visuals [131], they are limited in how many dimensions they can show [132] and may still lead to information overload [127]. Furthermore, previously proposed systems in this category have mostly been non-interactive which limit the capacity of the user to conduct exploratory analysis [133]. These systems fall short according to the Visual Information Seeking Mantra: Overview rst, Zoom and Filter then Details-on-Demand [134].

Works that combine both machine learning and data visualization methods are able to bypass some of the limitations seen in systems that only leverage one such methodology. Such systems are starting to appear [113]. However, most works leveraging both machine learning and data visualization methods have focused on cohort-level visualizations rather than patient-level visualizations [107–112]. Furthermore, like for data-driven Infobuttons, few of these systems have been evaluated for usefulness or usability at the point of care. Methods outside of the health domain that can inform future research include automatic visual summaries of temporal new stories and topic modeling [135, 136].

2.1.3 Alert clinical decision support

Alert clinical decision support produces a focused output such as a prediction about a speci c outcome, an alert, or a set of recommendations to clinicians in order to help direct action and prevent medical errors in the context of patient data.

Type of methods for Alert CDS

Of the three clinical decision support types, alert CDS has the most sustained interest in the literature. Early work on these systems date back to the late 1950's and continued with a recent surge. Similar to the Infobutton systems, most mature systems implemented and used by clinicians today leverage knowledge sources and expert curated rules [137–141]. Heuristics-based CDS have largely underutilized visualization techniques. Existing examples of the use of visualization showcase patient data alongside the alert or recommendation [142, 143]. Other work have proposed the use of visualization for knowledge base maintenance at the back-end of alert systems but not for the use of clinicians [142].

By contrast, the bulk of recent published work has focused on developing machine-learning methods that have the potential to assist in future alert CDS. Proposed machine learning methods have tackled a wide range of CDS applications and have leveraged a diverse set of approaches (Figure 2.3). Applications of machine learning methods developed for use in future alert CDS systems comprise disease and disease-stage prediction, optimal treatment prediction, and readmission and mortality prediction. The most popular machine learning approaches explored in recent years include deep learning and probabilistic methods.

Only a few systems leverage both machine learning and visualization. Systems that do utilize both methods motivate the use of visualization for added interpretability, model transparency, data provenance, and usability [110, 118, 144–146].

Gaps and opportunities for ML and dataVis in Alert CDS

Heuristics-based alert CDS have been found to improve healthcare processes, but that there is still little robust evidence of leading to improvements in clinical outcomes, costs, workload and ef ciencies [242, 243]. Commonly cited limitations of heuristics-based alert systems pertain to their narrow clinical focus, most likely due to the need for manual curation of clinical expertise in the systems. Few systems are `high-throughput' or able to assist on wide range of conditions and patient types. In practice, this can translate in multiple CDS systems, each relevant to a speci c

Table 2.3: Examples of Alert CDS by method type

Papers	Type of method utilized	Description
Warner et al. 1972, Warner 1979; Kuperman et al. 1991 [137, 141, 235]	Heuristics	Rule-based logical operators to assist with diagnosis
Miller et al. 1982; 1989 [140, 236]	Heuristics	Knowledge-based system that can construct and resolve differential diagnoses. Evaluated for accuracy compared to human experts. Evaluated for clinical utility
Goldstein et al. 2000; Gennari et al. 2003 [142, 143]	Heuristics+ dataVis	Guidelines and ontology-based treatment recommendation system for chronic disease. presents the patients raw data related to the chronic problem such as the patient's blood pressure readings over time
Warner et al. 1964 [150]	ML	Use Bayes' Theorem to the diagnosis of congenital heart disease. Compared accuracy of system to that of clinical experts
Wang et al. 2014 [190]	ML	Use unsupervised probabilistic model to model disease progression.
Tsoukalas et al. 2015 [118]	ML+dataVis	Partially observable markov decision process model. interactive graphical interface for optimal treatment for Sepsis. Includes visual of patient vital history over time, state transition probabilities, patient state history, and optimal action. Evaluate generalized error of approach and in external tasks of mortality prediction and length of stay prediction
Jeffery et al. 2017 [144]	ML+dataVis	Mobile app to showcase the predicted probability of cardiac arrest overtime, including forecasted risk for the next 24 hours. Evaluate tool for usability in a lab setting with target audience
More papers by method category:	Heuristics	[138, 139]
	ML	[26, 97, 99–101, 103, 106, 120, 121, 129, 130, 149, 151–155, 158–170, 172–189, 191–197, 199–201, 205–212, 214–223, 227, 229, 230, 237–240]
	ML+dataVis	[145, 146, 241]

Figure 2.3: Machine learning applications and approaches for alert CDS. Applications include disease classi cation or prediction [26, 97, 99, 100, 102, 106, 121, 129, 147–185], disease progression [101, 105, 118, 164, 186–200] , hospital readmission [201, 202], mortality prediction [116, 119, 203, 204], treatment-response prediction [103, 111, 130, 205–210], treatment recommendation [209–215], treatment identi cation [118, 216–220], and intervention prediction [209, 219–223]. Approaches include probabilistic methods [97, 150, 156, 162, 164, 178, 185, 193, 199, 200, 216, 222, 224–228], deep learning [100–103, 152, 156, 171, 173, 180, 181, 184, 189, 191, 195, 198, 218, 219, 229–231], support vectors [156, 163, 169, 178, 183, 189, 191], regression [151, 158, 169, 172, 175, 194], decision trees [166, 169], collaborative ltering [187], clustering [206, 232], reinforcement learning [59, 220, 233], and outlier detection [234].

subset of patients, with a need to deploy and manage them each to support diverse types of patients and clinical contexts. This can lead to `alert overload', with too many systems ring alerts to clinicians, each with little awareness of the others.

Adding data visualizations to heuristics-based alert CDS can help with interpretability and data provenance, leading to higher con dence in the system and usability. However very few works have explored this research space.

Introducing machine-learning techniques into alert CDS can help generate evidence directly from gathered clinical data, reducing the need for clinical knowledge to be codi ed manually by experts [129, 153, 159]. Moreover, machine learning methods can also handle many more predictors and complex relationships such as non-linearity, interactions, and temporality that would be hard to codify in knowledge-based systems [159, 192, 244]. Machine learning methods can also handle data with missingness, sparsity, noise, and irregular sampling [245, 246]. However,

machine learning methods intended for CDS have often been criticized as uninterpretable, prone to data biases, and dependent on the data they are evaluated on [26, 247–249]. This can make the comparison of models problematic when evaluated on different data and also be regarded as `too risky' to incorporate into clinical decision making. Other signi cant limitations of data-driven alert CDS is their lack of alignment with clinical work ows, with few proposed methods evaluating clinical utility with clinically meaningful metrics, and they have not been deployed to clinical settings [25, 250]. For instance, some approaches which ignore when data are generated in the clinical work ow, can lead to data leakage when predicting outcomes and would not be possible to implement.

While alert CDS that introduce data visualization for the end user are often more mindful of user-centered considerations such as the clinical work ow they attempt to support, they too have largely been evaluated on model accuracy, face validity of visualization, and interface usability in a laboratory setting [118, 144, 146, 241, 251]. The need for more interpretable and transparent learning methods has also been recognized outside of the health domain. Several reports have cited the integration of data visualization for the interpretation and understanding of machine learning methods and their results as key [252, 253].

Chapter 3: Iterative user-centered design approach for longitudinal patient record summarization

Involvement of end users in the design process in many domains has been widely considered to be a key factor for product usefulness and usability [254]. One such design methods that puts the intended user of a product at it's center is called `User-Centered Design' (UCD). The approach originates with seminal work by Donald and Draper [16] from the mid 80's. UCD approach is meant to overcome limitation of the system-centered design paradigm, where system designers focus on what the system should look like without much consideration of the intended users, and insist that users learn the system, rather than the system adapt to its users. UCD process requires the treatment of users as the subject of study and follows early focus on observing and understanding user and tasks in design, empirical evaluation and measurement of user interactions, and iterative design processes that involve cycling through design, evaluation, and re-design [16].

This dissertation follows an iterative UCD approach for collecting and iterating through the design speci cations of a longitudinal summarization system (steps (1) and (2) of gure 3.1). This chapter describes the needs nding collection process (step (1) in gure 3.1) and the initial round of design, low delity prototyping, and testing (step (2) in gure 3.1). The chapter ends with our translation of the collected information needs to a set of design requirements. Another round of the iterative design process (further prototype development, heuristic evaluation with target users, and usability study) is described in Chapter 5.

3.1 User-centered design of clinical decision support systems

UCD approaches have also been utilized (with varying degrees) in the health domain, including in the design of electronic health records [255], patient facing technologies [256], and consumer

Figure 3.1: Diagram of design process. Figure adapted from [12].

mobile health applications [257]. In fact the Of ce of the National Coordinator of Health Information Technology (ONC) has included UCD requirements for certi cation criteria of EHRs [258].

The increased role of users in the design process have been shown to be effective in designing and re ning health IT system, signi cantly reducing usability problems [259–261] and providing a low cost method for early detection of system error [262]. A recent literature review identified 24 studies leveraging UCD for health-related interventions, 9 of which were provider facing applications, 11 were patient facing, 2 were both, and 2 were for care-givers. Of the 9 provider facing interventions, 6 were found to be successful (de ned as improving all tracked metrics) and 3 studies with mixed results [12].

3.1.1 User-centered design of patient record summarization systems

Under the UCD approach, the design process begins by gaining an understanding of the information needs and work ow of target users (Figure 3.1 (1)). The patient summarization we are pursing is meant to assist clinicians in understanding the patient case when performing patient chart review. Patient chart review can be done in different points of the clinician work ow, which may

require different amounts of detail and time commitment. One use scenario of the patient summary is at the point of care, where clinician needs a gist of the medical history. Another scenario of use is during admission time, when clinicians needs an in-depth history of the patient case. Finally, clinicians and researchers also perform chart review in non-patient facing work ows in clinical research settings.

Studies assessing the information needs of clinicians that conform to the UCD approach have followed various methods (and a combination of these methods). Methods have included literature review [263–266], expert consultation [263–265], semi-structured interviews with clinicians [113, 267, 268], ethnographic studies [266–271], simulated recall [272], focus groups [265, 266, 271, 273].

In our work we collect the information needs of clinicians from patient longitudinal summaries by reviewing the relevant literature describing clinician information needs when acquiring a mental model of the patient case. To con rm the identi ed informational needs from the literature are consistent with the informational needs of clinicians seeing patients dealing with chronic disease and multi-morbidity we triangulate our ndings with the our selected use case, clinicians treating HIV patients. We construct a low delity prototype bases on the information needs we identify and run a formative study with clinicians to discover additional information needs and re ne our design requirements. In Chapter 5 of this dissertation we describe additional information needs gathering after heuristic evaluation with experts.

Previous works have built a substantial body of knowledge detailing the goals and processes by which clinicians obtain an overview of the patient case. These goals help us identify what are the information needs of clinicians that stem from these tasks.

Identi cation of patient problemsThe primary goal of clinicians when trying to get an overview of a patient case is the identi cation of the patient main problems [274]. This goal is often approached through a review of recent clinical notes [274, 275]. The required degree of problem comprehensiveness has been found to vary from complete comprehensiveness to a suf cient overview of patient problems that is enough for them to act on [275]. Different degrees of compre-

hensiveness is related to the time constraints the clinician has when obtaining the patient overview, which requires prioritization of problems.

Validation of identi ed patient problems. After identifying patient problems from one source of the patient record, clinicians often use other data sources from the patient record as well as historical data to validate the problems they previously identi ed for the patient [6, 274, 276]. This is sometimes done through directly locating speci c details related to the problem at hand such as medication or laboratory tests [113]. Moreover, throughout the patient data review clinicians are still on the lookout for problems they may have overlooked which is an indication that they are aware that no single source of data can be trusted to be fully complete [274].

Assessing problem status. Once clinicians have con dence in the patient problems, they aim to assess the status of the problem, as in whether it is an active or resolved problem and if it's an active problem then whether it's worsening, getting better, or stable. They do this by focusing on the temporal structure of data in the patient record, using older data in comparison to recent data. They are also on the lookout for change in the data such as medication orders that may signal a problem status change that prompted previous clinicians to change the patient's medications [6, 274].

Obtaining a temporal understanding of patient case. Clinicians aim to get historical overview of the patient case. This information is often sought when clinicians want to understanding the chronology of symptoms, developments, and the patient current state in more context [6, 113, 275].

Identifying correlations and relationships in the patient data. Clinicians look to identify correlation between the patient data points to form and test hypotheses regarding the patient state and reaction to treatments [6, 113, 275]. This has been especially emphasized when patient suffer from multimorbidty which complicates the patient case [6].

3.2 Clinicians' information needs at the point of chart review for patients with multimorbidty

As the information needs collected from the literature largely re ected information needs of general practitioners from longitudinal summaries we set out to assess whether they are consistent with the information needs of clinicians treating HIV patients. To this end we set out to assess whether a patient record summarization that showcases problems, their salience over time, and how they relate to one another is useful in the context of HIV patient care. We did so by performing semi-structured interviews with two physicians from the HIV clinic at NYPH about their perceived use for such a summary and their general informational needs about patient history. The interviews were designed to take about 30 minutes.

At the beginning of the interview we gave a general description of the proposed summarization system. Clinicians were then asked several questions regarding the preserved utility from such a summary. Questions included i) where in the clinical work ow the proposed summary would be useful? ii) what they would use it for? and iii) what is the hardest or most time consuming part of the process is now? Clinicians were then asked what they would want to know about the history of the patient that is currently hard to identify in the medical record. They were also asked how the care for HIV patients is different from the care of other patients and whether the proposed summary could support those unique needs. The clinicians interviewed were two Assistant Attendings working in NYPH's HIV clinic. An analysis of the interviews identi ed the following themes.

Summary perceived utility A longitudinal Summarization system assisting clinicians in identifying patient problems and there change over time would bene t clinicians when treating HIV patients. They noted that this type of summary would be especially bene cial in the HIV practice at NYPH since clinicians often need to see patients they are not familiar with and thus need to quickly understand the patient medical history. Moreover, these HIV patient suffer from many amounts of comorbidities and come to seek care on a regular basis and thus have complex medical histories with a lot of medical data documented in their records.

Identification of change over time. When asked about what features in a patient summarizer would be important they noted the intensity of the problem over time and any dramatic changes in the status of a problem could help them prioritize care. They also noted that patient hospitalizations were of note and the interventions they received during the hospitalization. In addition to their history of visits, which could clue clinicians in to adherence and the general state of the patient.

Uniqueness of HIV patients. The clinicians emphasized that comorbidities are a large factor in the care and the engagement of patients in their care. Thus there is an increased need to keep close watch on comorbidities as they can out pace issues caused by HIV such as cardiac problems and pulmonary problems. Moreover, problems associated with aging seem to arise sooner in the HIV patient population. Furthermore, patients are seen by other clinicians to monitor their chronic conditions and for acute problems and thus are hard to monitor.

Proposed scenarios of use for the summarization system. The clinicians indicated that there were multiple scenarios in which the patient summarizer could be of use. The most primary scenario of use mentioned was to aid in patient chart review in outpatient settings when clinician is first familiarized with the patient case or when the clinician needs a quick review of a previously seen patient, especially if the patient is treated by multiple clinics. The summary was also said to be helpful for patient chart review in an inpatient HIV services or for sub-specialists since patients are more likely to have never been seen by the caring clinician. Finally, the summarizer could also be used during patient consultation to facilitate joint decision making.

3.3 Iterative user-centered design of the summarization system

The next step in the UCD process is ideation of potential design solutions to the user information needs identified (Figure 3.1 (2)). We start with considering a visual design approach that would support showcasing patient problems and their change over time. Previous works have proposed temporal view of patient data but did not explicitly capture change in salience of problems in a single visual. Previous works on temporal visuals of patient data could visually convey change in salience by the density of observations assigned to a certain period [8, 80]. HARVEST showcased

a problem dominance in a time period, but required the user to scroll over time and observe the change in the font size of a word which could be dif cult to track [78].

3.3.1 Design ideation

To support the visual encoding of changes in problems over time we were inspired by several works outside of the health domain such as representation of story lines summaries using 'metro maps' (Figure 3.2) [135] and visualization of themes changes over time in large corpora [277]. From the literature healthcare visualization literature we were inspired by works using Sankey diagrams for visualization of different care paths of patient cohort (Figure 3.3) [108, 109, 112].

Figure 3.2: Information representation through `metro maps of information'. Visual taken from original work [135].

3.3.2 Prototyping 1.0: Sankey diagram to represent patient problems over time

We leveraged the sankey diagram visual encoding to represent patient problems over time. To our knowledge our work represents the rst to propose such visual representation of patients over time. The preliminary design of the summary is presented in Figure 3.4. The x-axis of the visual shows three selected time slices of the patient's record. The problem proportion identi ed

Figure 3.3: Cohort visualization of treatments and outcomes. Visual taken from original work [108].

in each time slice is represented by a gray rectangular node. Problems with higher proportion are represented with longer nodes. Nodes relating to the same problem over time are connected using a colored link to help the user track the same problem over time. The positioning of the nodes on the y-axis have no semantic meaning. In the rst prototype of the tool, the problem proportions in each time period were obtained using an off the shelf Correlated Topic Model (CTM) [278]. The CTM was used to summarize the patient record only using a single data type, diagnosis codes. The CTM in its original form can only handle a single data source at a time, and often is used to identify themes in large sets of written documents.

We implemented the summarizer prototype as a web-based interactive tool using the Javascript library D3.js. Users interact with the tool by entering the patient id number they wish to summarize, the number of top problems to view, the criteria by which top problems are selected, and whether the problems are grouped or visualized individually. The user can also select to view automated labels for the problems assigned using the most probably diagnosis code in the topic. Hovering over each problem link shows the diagnosis code cloud of the phenotype learned by the model.

Figure 3.4: Preliminary design of summary of patient problems over time using Sankey diagram

3.3.3 Testing 1.0: Formative usability study

To assess the usability of the proposed visualization and its ability to support clinicians in identifying patient problems and their change over time we conducted a formative user study. The user study was performed with two physicians (participant 1 and 2), two patient cases of similar complexity (patient A and B), under two study conditions (Condition 1 and 2). Patient data was automatically summarized using computational phenotyping model that learns problem correlations but only uses a single data type, patients' diagnosis codes. Under Condition 1 a participant starts by constructing a problem list using the visualization summary of the patient provided by the tool, then the participant moves to the patient record in the EHR system to validate or change the problem list. In Condition 2, the participant constructs the patient problem list using the EHR system alone.

Participant 1 reviewed patient A under Condition 1 and the patient B under Condition 2. Participant 2 reviewed the patients under the opposite conditions (patient A under Condition 2 and patient B under Condition 1). Since participants have never seen the tool before, a short overview of the tool and its functionalities is given to each participant. A gold standard problem list is generated

by a third clinical expert. The amount of time participant took to complete the patient problem list under each condition was tracked. The problem list generated under each condition is compared to the gold standard list for each patient. Precision and recall are calculated and used to compare the study conditions. An error analysis of the tool is preformed to identify if miss-representations of the patient problems are caused by the computational model, visualization decisions, or the underlying data. At the end of the task each participant is asked for their feedback of the tool.

An analysis of study results indicated that the problem lists from Condition 1 (using the EHR+ tool) were more complete but took longer to construct (an average of 12 minutes versus 6 minutes under Condition 2 (EHR only)). Errors in the visualization tool were largely due to incomplete diagnosis code data in the patient record, fewer errors were due to modeling and visualization decisions. The study ndings suggested that results could be improved by modeling additional data types other than diagnosis codes codes such as clinical notes, laboratory tests, and medications. This suggests that the investigation of a more advance inference method that allows for multiple input types is warranted.

Feedback from participants also indicated the desirability of patient speci c problem labeling and word-clouds. Furthermore, the study of usefulness of the tool with varying degree of record complexity is required. The results of the pilot study were presented in the 2017 Visual Analytics in Health Care (VAHC) workshop.

3.4 Design requirements of the summarization system

To generate the design requirements of the patient longitudinal summarizer we translated the gathered informational needs of clinicians that we collected from the literature review, the clinician interviews, and the formative study of initial summarizer prototype. In Table 3.1 we list the information need and its corresponding design requirement. We operationalize each design requirement using i) visualization; ii) machine learning; and iii) interactivity or a combination of methods. We describe the selected operationalization of each design requirement below.

Table 3.1: Information needs to design requirements

Information need	Design requirement
1. Key problem identi cation	Problem-oriented summary that generalizes to different patient complexities and problems and is able to perform effective content selection to identify patient's top problems
2. Multimorbidity summarization	Conveys patient's problems in context of each other; allows for identi cation of correlations among data and problems
3. Problem veri cation	Allows access to patient data related to a problem to verify problem and establish trust in system
4. Problem Status ascertainment	Captures salience of a given problem and any changes to problem status
5. Longitudinal information regarding patient problems	Temporal view of problems
6. Reduction of patient data information overload	Generates a digestible summary of patient data by providing meaningful information reduction

R1: Supportproblem oriented view of the patient record. This aims to assist users identify patients' main problems. The design requirements is operationalized through robust inference of patient problems based on their clinical documentation. We elect to use a computational phenotyping model we develop that is able to perform inference over structured and unstructured data from the patient record. The model simultaneously identi es a list of patient problems and is able to prioritize problems by their respective salience in the record. The model is unsupervised and is able to handle different patient complexities (with little or many problems) and is able identify various types of problems. The development and evaluation of the model is discussed in Chapter 4 of this dissertation.

The visual and interactivity of the summary interface is the second way that the design requirement is operationalized. The visualization we elected to use shows all or selected patient problems. We do so through the use of sankey diagram, where each sankey link represents a problem. Through the system's interactivity, users can select to view all of the patient's top problem and zoom-in and view a single problem at a time.

R2: Supportsummarization of patients with multimorbidty. The model performs simultaneous inference of patient problems and accounts for correlations between problems. The model also explicitly learns relationships between problems it identi es from a cohort of similar patients. Problem relatedness is used in the summary interface and allows users to elect to view problems and their related problems at the same time.

R3: Support users inverifying patient problems. Allow veri cation of patient problems by checking associated patient data from different sources in the patient record. The design requirement is execute by leveraging an interpretable phenotyping model that allows users to identify the patient underlying data and the problem they were associated with. This allows for users to verify suspected problems but also apply judgment regarding the accuracy and trustworthiness of the summarization. The visual design of the tool needs to allows this information to accessible and easy to navigate.

R4: supportproblem status ascertainmentof each identi ed problem. Problem salience is identi ed by the model through the amount of the patient data attributed to each problem in each time period. Problem salience is visually encoded in each time period through the width of the sankey link associated with the problem. Change in salience is visually encoded through the change in width of the sankey diagram.

R5: The system must supportemporal viewsof patient problems over time. To support such a need the ML model used to summarize the patient problems needs to be able to provide a temporal summary of patient and the visualization needs to support the visual encoding of such a summary.

R6: Supportinformation overload reduction through providing interpretable dimensionality reduction of patient data. This is achieved through digestible abstractive summarization of patient raw data and grouping into clinically meaningful problems. Further problem abstraction is provided by allowing users to simultaneously view related problems. The visual interface supports

this design requirements by providing a high level overview of the patient problems through the sankey diagram. Users are able to leverage the summary interactivity to zoom-in- and out of the patient underlying data.

Chapter 4: Summary and content selection through joint phenotype learning

4.1 Introduction

The EHR have improved the availability of patient records, but this has not always translated to increased availability of relevant information to clinicians [279]. This is partly because increased amounts of data in EHRs has made it more dif cult for clinicians to review patients' previous medical histories and obtain an overview of the patient record [275]. Increased amounts patient data have also raised concerns regarding clinician information overload [5], having effects on care quality [280], and patient safety [281].

Patient record summarization has been suggested as a valuable tool to support clinicians in making sense of increasingly large patient records[282]. There are a number of open challenges associated with robust summarization of clinical documentation [128], including content selection— identifying the right summary elements at the right granularity in the input patient record— and content organization—organizing summary output in a coherent and actionable fashion for the clinicians, all the while preserving data provenance.

Previous work has shown that problem-oriented summaries support the information needs of clinicians [7, 73, 78, 283]. High-throughput computational phenotyping methods that utilize both structure and unstructured EHR data are attractive for identifying a patient's problems in a robust and scalable fashion [93]. Other unsupervised methods to identifying phenotypes using multiple data types such as matrix and tensor factorization have been shown to generate interpretable phenotypes but have largely been restricted to the use of structured data [129, 227, 284, 285] or clinical notes in isolation [116]. Considering the characteristics of EHR data (missingness, heterogeneity, uncertainty), Bayesian generative approaches are attractive to handle them and provide easily interpretable outputs that quantify their uncertainty.

EHR-driven phenotyping more generally have been proposed for several use cases, most often for cohort identi cation [286–288], patient population characterization [289, 290], but also for clinical decision support [61, 78, 291]. Approaches for EHR phenotyping divide into those leveraging expert-curated rule based algorithms [292] and those utilizing computational algorithms [93]. Computational algorithms proposed for EHR phenotyping include methods that are fully unsupervised [93], semi-supervised [293], and supervised [294, 295]. Strategies for feature engineering also ranged from expert feature engineering [296], automated feature selection [293, 297], to representation learning [99, 102].

To enable a problem-oriented summarizer to identify a target patient's comprehensive list of problems and their salience, we propose a probabilistic machine learning approach that can identify a large number of phenotypes/problems using patients' structured and unstructured data in an unsupervised fashion. The machine learning model is trained on the EHR data of many patients to simultaneously learn probabilistic de nitions of many phenotypes at the same time. Figure 4.1 shows a graphical schema of the proposed approach .

To identify the appropriate granularity of the learned phenotypes, the model is trained on a target patient population of the same clinic. Each phenotype de nition is composed of diagnoses, medications, laboratory tests, and clinical notes that have been observed to commonly co-occur in the training patient population. Figure 4.2 shows an example phenotype learned by the model. Figure 4.3 shows phenotype-phenotype correlations learned by the model. Phenotypes are labeled with their most probable diagnosis code. The learned phenotypes from the model are then used to summarize a single patient EHR data over time. Figure 4.4 shows an automatically generated example summary of a single patient record over a ve year period that leverages the proposed approach.

Since our use case for summarization are HIV patients, we focus on phenotyping the patient population from the HIV clinic at NYPH. We hypothesize that (1) the model will learn many clinically valid phenotypes and phenotype relationships; (2) training the model on an HIV-positive patient population will result in the identi cation of several HIV phenotypes, representing the

Figure 4.1: Schema of summary and content selection approach using joint phenotyping.

different presentations and progression stages of HIV—a granularity that would likely be missed if trained on a more general and heterogeneous patient population; 3) the model will also learn non-HIV phenotypes, representative of the many comorbidities of HIV; 4) the model will identify correlations among phenotypes that indicate clinically valid relations of different types beyond simple is-a relationships.

4.2 Methods

4.2.1 The model

The model we proposed is based on the correlated topic model (CTM) [278]. In our context, topics are equivalent to phenotypes and documents are the patient records. We make a methodological contribution by expanding the CTM and its inference method to support multiple input sources, beyond the single input source usually assumed in topic modeling. We make this important expansion to the model since unlike topic identi cation in general text, clinical documentation is more than just clinical notes. Instead, our model is able to learn phenotype de nitions through identifying co-occurring patterns in clinical notes, laboratory tests, ordered medications, and diag-

Figure 4.2: Example of learned phenotype and its probabilistic de nition across the four data types (yellow for diagnosis codes, green for notes, purple for medications, and blue for laboratory tests). The mostly likely diagnosis code is assigned as label for the phenotype.)

nosis codes. Incorporating multiple sources of data into the phenotypes de nitions allows for more robust phenotype de nitions that can help overcome the inaccuracies present when just relying on an single source of patient data (e.g., diagnosis codes). This is supported by previous work that has shown that incorporating heterogeneous data yields superior phenotypes [93].

It has been previously proposed to leverage topic-model like models to learn clinical phenotypes. Our model differs in that we do not assume that phenotypes identi ed in each patient record are independent from one another. We remove the assumption of independence by allowing for phenotypes to be correlated. To do this our model, like the original CTM, replaces the traditional Dirichlet distribution used in Latent Dirichlet Allocation (LDA) [298] to govern topic proportions with a logit-normal distribution [278]. The logit-normal distribution allows for phenotype proportions in each patient record to be correlated with one another (through the normal covariance

Figure 4.3: Example of ve learned phenotypes and their learned correlations ( ).

matrix) but also add up to 1 or 100% of the patient record, as desired when modeling proportions. Changing the previously assumed Dirichlet distribution with a logit-normal distribution removes the conditional conjugacy between the posterior distribution and prior distribution of the phenotype proportions. To perform posterior inference Wang and Blei [299] propose Laplace Variational Inference, a generalized form of variational inference (VI) that can handle non-conjugate models. However Laplace VI previously only supported a single data types as input. In this paper we generalize the proposed Laplace VI even further to allow for multiple input types. This makes the model inference especially relevant to clinical data which contains many different data types. The model training is time-agnostic and treats each patient record as bag of observations, one for each data type. While motivation behind the model is to assign phenotypes on a single patient level for patient-level summarization, in this paper we focus on the learned phenotypes on the population level. Each phenotype is labeled using the most probably diagnosis code.

The generative process of each patient record (D) with number of tokens for data types

Figure 4.4: Example of patient-speci c summary over ve years. The top ve most salient problems in 2019 are selected and visualized to showcase how their documentation has evolved through time. In this setup, the summary was produced at the year level by binning the patient's documentation for that time resolution. The patient has HIV-speci c problems, although their HIV is becoming asymptomatic, as well as comorbidities, all cardiac in nature. (Relations among the inferred phenotypes are not shown. Dates are changed to maintain patient privacy.)

is provided below. The graphical representation of the model is presented in Figure 4.5.

1. Draw log phenotype proportion α: $\alpha \sim \mathcal{N}(\mu_0, \Sigma_0)$
2. For each token $(w_{d,n})$ in data type d $(d = 1, \ldots, D)$:
   (a) Draw phenotype assignment $z_{d,n} \mid \alpha \sim \text{Cat}(\pi(\alpha))$
   (b) Draw token $w_{d,n} \mid z_{d,n}, \beta_{1:K} \sim \text{Cat}(\beta_{z_{d,n}})$

4.2.2 Probabilistic inference

The phenotype de nition and their correlations with one another are obtained through performing Bayesian posterior inference which estimates the conditional probability of the unobserved or latent model variables given the observed model variables. In the case of the proposed model

Figure 4.5: The graphical representation of the multi-input correlated topic model. Multiple inputs are represented by the additional plate notation M that is not present in the single-input CTM model.

this means calculating the probability of the phenotype proportions of each patient record and (phenotype assignment of each input given the observed patient data ( ) and phenotype distributions ( ), or ? ¹ a – ∫ G – ∇. When the posterior distribution has a conjugate prior this greatly simpli es the Bayesian analysis and allows for the use of popular sampling methods for approximate inference such as Markov chain Monte Carlo sampling such as Gibbs sampling as employed in [93].

However, conjugacy limits the types of distributions used in the model, and thus restricts the exibility of data modeling. In order to allow for phenotypes to be correlated with one another the prior distribution used to model the phenotype proportions in the patient record needs to allow for phenotype correlations. Since that is not possible with the Dirichlet distribution, it needs to be replaced with a different distribution that meets this criteria. However since the Dirichlet distribution is the conjugate prior to the multivariate distribution used to model the phenotype data assignments, this modeling change means that the model losses its conditional conjugacy. Hence, deterministic approximate inference methods such as VI is more feasible than other sampling methods.

By contrast to sampling approximation methods for inference, the theoretical guaranties of convergence of VI methods to the true posterior have been less studied. However, VI has become a popular inference method in Bayesian statistics as it tends to be faster and scale better with large

and complex data [300]. Even in VI, some popular implementations such as mean- eld variational depend of conjugate models. Wang and Blei [299] propose Laplace VI, a generalized form of VI that can handle non-conjugate models. The method uses Laplace approximations in the coordinate ascent updates within the variational optimization problem. This methods was shown to generalize to different types of non-conjugate model and have superior performance compared to the original ad-hoc inference method previously proposed here [278]. In this paper we generalized the Laplace VI for multiple input types. The mathematical derivation of the Laplace VI with multiple input types is shown below.

As presented in the graphical model (see Figure 4.5), the under-script represents the -th input type, where = 1– •••– ". The derivation below contributes to the previously proposed inference by [299] by allowing for M input types instead of a single input type. The model is represented by the joint probability distribution in equation (1). The inference problem is to solve for the posterior distribution which is the conditional distribution of the latent variables and given in equation (2).

$$?^1a– I–^oG= \ddot{O}_{<=1} ?^1G< jI_<{}^o?^1I_< ja^o?^1a^o \tag{4.1}$$

$$?^1a– JG^o = \frac{– ?^1a– I–^oG}{?^1a– I–^oG3Ia} \tag{4.2}$$

The integral in the denominator of equation (2) is intractable to compute exactly [278]. As proposed by [299] the posterior is approximated using Laplace VI through optimization. A family of densities are posited over the latent variables. The model assumptions include:

1. The variational distribution is fully factorized:

$$@^a– I^o = @^aa^o \ddot{O}_{<=1} @^aI_<{}^o \tag{4.3}$$

2. a is real valued and ?^1a^o is twice differentiable with respect to a

3. The distribution $?^1I_< ja^o$ is in the exponential family:

$$?^1I_< ja^o = \ ^1I_<{}^o 4G\ ?[\ ^1a^{o)}\ CI_<{}^o \quad 0^1[\ ^1a^{oo}g \tag{4.4}$$

4. The distribution $?^1G_< jI_<{}^o$ is in the exponential exponential family such that:

$$?^1G_< jI_<{}^o = \ ^1G_<{}^o 4G\ ?CI_<{}^{o)}\ \ddot{Y}\ CG_<{}^o{-}1\ ¡\ g \tag{4.5}$$

In VI the approximation for the posterior distribution is obtained through minimizing the Kullback-Leiber (KL) divergence to the exact posterior.

$$@^1a{-}\ ? = 0A6<8= \ !^1\&^1a{-}\ ?jj\%^1a{-}\ JG^o \tag{4.6}$$

Under standard VI theory minimizing the KL divergence between $@^1a{-}\ ?$ and the true posterior $?^1a{-}\ JG^o$ is the same as maximizing the lower bound of the log marginal likelihood of observed data. Using Jensens's inequality the variational objective $!^1@^o$ is de ned by equation (5).

$$\begin{aligned} ;>6\ ?^1G^o &= ;>6 \int^1 ?^1a{-}\ I{-}\ ^oG3I3a \\ &@^{»};>6^1\ ?^1a{-}\ I{-}\ ^oG^1¼ \quad @^{»};>6^1@^1a{-}\ ?^o1¼ \\ &= \quad @^{»};>6^1\ ?^1a{-}\ I{-}\ ^oG^1¼ \quad @^{»};>6^1@^1a^o \quad \ddot{O}_{<=1} @^1I_<{}^o1¼ \\ &! \ ^1@^o \end{aligned} \tag{4.7}$$

Setting the partial derivative of $!^1@^o$ with respect to $@$ to zero provides the optimal variational updates to $@^1a^o$ and $@^1I_<{}^o$ seen in Equations (9) and (10). Where $?^1a^o$ is conjugate to $?^1I_< ja^o$ then equations (5) and (6) have closed form solutions. In the case of this non-conjugate model [299] put forward approximates to the updates using Laplace approximation.

$$@^1a^o\ /\ 4G\ ?\ @I^o{}_{»};>6 \ddot{O}_{<=1} ?^1I_< ja^o ?^1a^o1¼g \tag{4.8}$$

@¹I₁º / 4G‹P @aº»;>6 ?¹G₁jI₁º?¹I₁jaº¼g

⋮ (4.9)

@¹I<º / 4G‹P @aº»;>6 ?¹G< jI₁º?¹I< jaº¼g

The following is the derivation of the variational update @¹aº using the previously stated assumption that ?¹I< jaº is assumed to belong to the exponential family.

@¹aº / 4G‹P @Iº»;>6 Ö<=1 ?¹I< jaº?¹aº¼g

= 4G‹P @Iº»;>6 ?¹aº ¸ Õ<=1 ;>6¹I< jaº¼g

= 4G‹P @Iº» Õ<=1 ;>6¹ ¹I<º4G‹P[¹aº) C¹I<º 0¹[¹aººgºº ¸ ;>6 ?¹aº¼g (4.10)

= 4G‹P @Iº» Õ<=1 ¹[¹aº) C¹I<º 0¹[¹aººº ¸ ;>6 ?¹aº¼g

= 4G‹P @Iº 5¹aºg

The function 5¹aº in Equation (10) has no closed form and this is approximated with the following 2nd order Taylor approximation around â which is the a that maximizes 5¹aº.

5¹aº 5¹âº ¸ r 5¹âº¹a âº ¸ 1/2 ¹a âº) r² 5¹âº¹a âº (4.11)

Thus the update for @¹aº is approximate with N¹â, –r² 5¹âº⁻¹º

The suf cient statistics of the exponential family are:

$$\begin{aligned} {}^{1}I_{<}{}^{o} &= 1 \\ C^{1}I_{<}{}^{o} &= \tilde{O}_{=} I_{<=} \\ [\,{}^{1}a^{o} &= a \quad ;>6f \quad \tilde{O}_{:} 4G?^{1}ag \\ 0^{1}[\,{}^{1}a^{oo} &= 0 \end{aligned} \tag{4.12}$$

Using the suf cient statistics above5¹aº is the following:

$$\begin{aligned} 5^{1}a^{o} &= \tilde{O}_{<=1} {}^{1}[\,{}^{1}a^{o)} \quad 0^{1}[\,{}^{1}a^{oo} \quad \frac{1}{2}{}^{1}a \quad `_{0}{}^{o)} \quad {}_{0}{}^{1}{}^{1}a \quad `_{0}{}^{o} \\ &= [\,{}^{1}a^{o)} \tilde{O}_{<=1} f \quad @I^{o}\,»C^{1}I_{<}{}^{o}¼g \; \frac{1}{2}{}^{1}a \quad `_{0}{}^{o)} \quad {}_{0}{}^{1}{}^{1}a \quad `_{0}{}^{o} \end{aligned} \tag{4.13}$$

The rst derivative and second derivative of5¹aº are the following:

$$\begin{aligned} r\;5^{1}a^{o} &= c_{8}{}^{1}1_{»8=9¼} \quad c_{9}{}^{o} \tilde{O}_{<=1} \quad @I^{o}\,»C^{1}I_{<}{}^{o}¼ \quad {}_{0}{}^{1}{}^{1}a \quad `_{0}{}^{o} \\ &= \tilde{O}_{<=1} \quad @I^{o}\,»C^{1}I_{<}¼ \; c \quad \tilde{O}_{:=1} \; \tilde{O}_{<=1} \quad @I^{o}\,»C^{1}I_{<}{}^{o}¼¼ \quad {}_{0}{}^{1}{}^{1}a \quad `_{0}{}^{o} \end{aligned} \tag{4.14}$$

wherec / 4G?[¹aºg

$$r^{2}\,5^{1}a^{o}{}_{89} = {}^{1} \; c_{8}1_{8=9}, \; c_{8}c_{9}{}^{o} \tilde{O}_{:=1} \; \tilde{O}_{<=1} \quad @I^{o}\,»C^{1}I^{o}¼¼{}^{1} \quad {}_{0}{}^{1}{}^{o}{}_{89} \tag{4.15}$$

The update to@²I_<ºwhere< = 1–•••–ïs the following:

$$\begin{aligned} @^{1}I_{<}{}^{o} / \; 4G?^{1} \quad @a^{o}\,»;>6?^{1}G_{<}\,jI_{<}{}^{o}?^{1}I_{<}\,ja^{o}¼g \\ = 4G?^{1};>6?^{1}G_{<}\,jI_{<}{}^{o} \; , \quad @a^{o}\,»;>6?^{1}I_{<}\,ja^{o}¼g \end{aligned} \tag{4.16}$$

Using the exponential form of $\Psi(\theta_{<}|a^{o})$ and $\Psi(G_{<}|\theta_{<}^{o})$:

$$
\begin{aligned}
\mathrm{E}[\theta_{<}|\mathbf{w}_{<}^{o}] &= \mathrm{E}[\Psi(G_{<}|\theta_{<}^{o}), \mathrm{E}[\Psi(\theta_{<}|a^{o})]] \\
&= \mathrm{E}[\Psi(G_{<}|\theta_{<}^{o}), \mathrm{E}[\theta_{<}^{o}, \mathrm{E}[\log(\theta_{<}^{o})]]] \\
&= \mathrm{E}[\theta_{<}^{o}, \Psi(\gamma_{<}^{o}) - \Psi(\Gamma_{<}^{o})] + \mathrm{E}[\theta_{<}^{o}, \\
&\quad \mathrm{E}[\log(\theta_{<}^{o})]]
\end{aligned}
\tag{4.17}
$$

$$
\mathrm{E}[\theta_{<}^{o}] \propto \exp(\mathrm{E}[\log(\theta_{<}^{o})] + \mathrm{E}[\log(\Gamma_{<}^{o})]) \tag{4.18}
$$

4.2.3 Dataset

The model was trained on the EHR data of 7,523 patients from an HIV clinical from NYPH. The data spanned 8 years and included the data types: words from clinical notes, laboratory tests ordered, medication orders, and assigned diagnoses codes from across all clinical settings (inpatient, outpatient, emergency). For the purpose of the model training each patient record was restricted to the most recent 2.5 years data. The final training data set included the following total data counts and unique vocabulary size in brackets: total words from clinical notes: 128,034,516 (unique: 25,894); total laboratory tests: 463,524 (unique: 129); total medications: 510,820 (unique: 6,714); and total diagnosis codes: 246,623 (unique: 2,956).

4.2.4 Model training and parameter selection

The parameters of the normal distribution governing phenotype proportions were initialized with μ_0 equal to a zero vector and λ_0 set to the identity matrix. The phenotype distribution $\phi_{<}$ for each input type was initialized with a Uniform distribution over the (K-1) simplex. This equivalent to initializing topics with a Dirichlet distribution with parameterization of 1. A small amount of random positive noise was added to each uniform distribution so there was a small variation in the initial phenotypes. Three alternatives of the model were estimated (K=50, 100, 250).

To identify the best performing model of the three alternative number of phenotype (K=50,

100, 250), a clinical expert reviewed 20 randomly selected phenotypes from each model. The best performing model is further evaluated for the clinical correctness of the phenotypes and phenotype-relatedness learned by the model.

4.3 Evaluation Setup

We evaluate our hypotheses 1 through 4 using a mixture of qualitative and quantitative evaluations. Qualitative evaluation of the phenotypes and phenotype relatedness was performed by two clinical experts. The quantitative evaluation was performed through a comparison to the Clinical Classi cation Software (CCS), which provides expert-curated manual classi cation of diagnosis codes into largely clinically homogeneous groups [301]

4.3.1 Hypothesis 1: clinical validity

To evaluate the clinical validity of the learned phenotypes, 50 randomly selected phenotypes were evaluated independently by two clinicians. The phenotypes were evaluated according to their coherence, granularity, and label quality [93]. Previous works citing clinical evaluation of phenotypes by experts have reported the scoring of a single clinician [93, 129]. Since this scoring can very subjective, we opted for two clinicians to score the phenotypes and the nal score assigned is the average of the two clinicians. Since we did not want the opinion of one clinicians to be in uenced by the other, there was not adjudication stage in the scoring (common on qualitative rating tasks made by more than one reviewer). This made the qualitative evaluation a very stringent task. We provide an analysis of the agreement between the clinicians scoring which can illuminate the level of subjectivity of this type of evaluation.

Phenotype coherence

Phenotype coherence is meant to capture the quality of each learned phenotype according to its most probably observations. A coherent phenotype is de ned to describe a single condition with few or no unrelated observations (clinical words, laboratory tests, medications, and diagnosis

codes). The expert was asked to rate each phenotype as having: `bad coherence' (score=1) , `some coherence' (score=2), `good coherence' (score=3), or `excellent coherence' (score=4). Phenotypes with `bad coherence' should look like a random combination of observations, `some coherence' indicates the observations assigned to the phenotype are somewhat related to one another, `good coherence' indicates the phenotype is a very good representation of a disease, and `excellent coherence' indicates the phenotype de nition has almost no unrelated observations assigned to it.

Phenotype granularity

The clinical experts were asked to characterize the granularity of each randomly selected phenotype by assessing whether the model learned a `single disease' (score=3), a `group of diseases' (score=2), or a 'non-disease' phenotype (score=1).

Label quality

The representativeness of the automatically assigned phenotype label of the phenotype as a whole was evaluated. Each label was categorized by the clinical experts as `unrelated' to the rest of the phenotype (score=1), `related' to the rest of the phenotype (score=2), or `actionable' (score=3). Labels that were deemed as actionable are those representative of a single phenotype and have the appropriate granularity to provide a clinician information that could be used without additional information to guide further testing, diagnosis, or counseling.

Phenotype relatedness

Next, the clinical validity of the phenotypes-relatedness were evaluated by a single clinical expert. The expert reviewed all phenotypes relationships that were indicated to have a correlation greater than 0.5 correlation coef cient. Two sets of phenotype-relationships were evaluated: 1) positive phenotype relationships learned between "more common" non-HIV phenotypes, de ned as phenotypes that were represented in more than 5% of the patient population in our dataset;

and 2) positive relationships learned between "rarer" non-HIV phenotypes, represented in 5% of sample population or less. The justi cation for evaluating relationships between "more common" phenotypes is that the model ndings are grounded in more patient record, which could result in more robust ndings. However, evaluating phenotype relationships identi ed between "rarer" phenotypes could still be interesting to assess in case the model is able to identify less known clinical relationships.

4.3.2 Hypothesis 2: focus on HIV phenotypes

Our second hypothesis was that since HIV is a complex disorder with diverse presentations and severity among patients, the model would identify several distinct HIV phenotypes. In evaluating this hypothesis, we wished to understand to what extent the model is able to learn multiple clinically valid HIV phenotypes and also characterize what those phenotypes were. To do so we had an HIV clinical expert review all the phenotypes automatically labeled as `HIV'. The clinical expert was asked to i) indicate if the phenotype was clinically valid, ii) indicate if the phenotype was indeed an `HIV' phenotypes; and iii) give a more granular description of the phenotype if it was indeed an `HIV' phenotype in order to assess if the model identi ed disease progression, presentation, or acuity.

4.3.3 Hypothesis 3: focus on non-HIV phenotypes

To assess if the model was able to learn diverse phenotypes, representative of the many comorbidities of HIV we quantitatively compared the phenotypes learned to the disease groups identi ed in the CCS. We did this by categorizing all 250 learned phenotypes according to their labels' corresponding CCS level-1 category. If the model was able to learn phenotypes that t into many CCS categories, we would could conclude that the model was able to learn diverse types of phenotypes, beyond HIV.

4.3.4 Hypothesis 4: types of phenotype- relatedness

We performed two evaluations to assess whether the model was able to identify correlations among phenotypes that indicate clinically valid relations of different types beyond simple is-a relationships. The first evaluation included a clinical expert review phenotype-pairs identified by the model as highly related and determine what kind of relationship type the model learned. Example relationship types include comorbidities, same phenotype, phenotype sub-type, and others. In the second evaluation we counted how many significant phenotype-relations learned by the model indicated an is-a relationship, as evidenced by same level 1 CCS categories, versus a more diverse relation type such as comorbidity when spanning different CCS categories.

4.4 Results

The qualitative evaluation by the clinical expert indicated that the 250-phenotype model yielded the most coherent and granular phenotypes of the three models (K=50, 100, 250). All results below are described for the evaluation performed for the K=250 phenotype model.

4.4.1 Hypothesis 1: clinical validity

Phenotype quality

Of the 50 evaluated phenotypes from the 250-phenotype model, 10% of the phenotypes (n=5) were deemed to have no coherence (average coherence score of 1 or 1.5) while the large majority of evaluated phenotypes (n=45) were deemed to be coherent (with average coherence score of 2 or above) (see Figure 4.6). The most number of phenotypes were scored as having 'good coherence' (n=13), followed by 12 phenotypes with an average of 3.5 (between `good coherence' and `excellent coherence'). The `bad coherence' phenotypes were found to be non-disease specific, but instead captured documentation related to general primary care visits. Figure 4.7 shows the diagnosis codes of example phenotypes with coherence scores 1 (`bad coherence') through 4 (`excellent coherence') by both the clinical experts. The phenotypes in the example identified a clinic

visit phenotype (scored 1), grouping of cancers phenotype (scored 2), grouping of heart diseases phenotype (scored 3), and an Arterial brillation phenotype (scored 4).

Comparing the coherence phenotype scoring assigned by the two clinicians we found that the two clinicians had a low agreement on the exact coherence score assigned to the phenotypes (scores 1 through 4) but that the average difference between the scores was less than 1 point (0.9). This indicates that the clinicians evaluation of the phenotypes was not far apart. When comparing the clinician agreement on whether a phenotype was identi ed as not coherent (score of 1) versus coherent (score of 2 and above) the agreement was high, at 90% of the evaluated phenotypes (see Table 4.1). Of the 5 phenotypes that the reviewers did not agree on, 4 looked like HIV clinic well visits. The disagreement seemed to stem from whether the model identi ed a disease phenotype or a clinical-settings phenotype. An example such phenotype had the following top 5 diagnosis codes: `Human immune virus disease', `Obesity NOS', `Elevated blood pressure w/o hypertension', `Hypertension NOS', and `Laboratory exam NOS'.

Figure 4.6: Phenotype coherence scores. Average score across the two clinical expert scores. Score 1=`bad coherence', 2=`good coherence', 3= `very good coherence', 4=`excellent coherence'.

The phenotype granularity scores indicated that 90% of the evaluated phenotypes (n=47) had a granularity score 2 or greater (see Figure 4.8). This means that almost all of the evaluated phenotypes were deemed by both reviewers to identify a single or a group of diseases. The most number of phenotypes were assigned an average score of 2.5 (n=31), the next most prevalent score

Figure 4.7: Example phenotypes by coherence score assigned by the clinical experts. Each phenotype is represented here by its top diagnosis codes rather than all 4 data types for the sake of space. Score 1=`bad coherence', 2=`good coherence', 3= `very good coherence', 4=`excellent coherence'.

was 2 (n=13). This indicates that the model mostly identi ed phenotypes that were a group of diseases rather than a single disease. An example of a phenotype that had an average granularity score of 2.5 had different diagnoses codes identifying a fall or accident and different body parts such as shoulder, forearm, limb and hand. One reviewer scored the phenotype as identifying a single disease being `limb injury due to accident', while the other reviewer believed that since multiple body parts were identi ed the phenotype represented a group of diseases.

The phenotype labels were mostly found to be `related' with a score of 2.5 (n=21) or 2 (n=19) (see Figure 4.9). Only 3 phenotype labels were identi ed as `actionable' with a score of 3 by both reviewers. The feedback from the reviewers was that the diagnosis code used for the phenotypes

Table 4.1: Comparison of 2 clinician scoring for phenotype coherence

		Clinician 1	
		Not Coherent	Coherent
Clinician 2	Not Coherent	1	1
	Coherent	4	45

Figure 4.8: Phenotype granularity scores. Average score across the two clinical expert scores. Score 1='non disease', 2='group of diseases', 3= 'single disease'.

was too granular to adequately represent the entire phenotype.

Phenotype-relatedness quality

Of the learned phenotype-pair correlations, 471 (1.5% of all possible phenotype-phenotype pairs) were signi cant (correlation coef cient above 0.5 in absolute value). Of the 471 signi - cantly correlated phenotype-pairs, 395 where positive correlated (Figure 4.10) and 76 were neg- atively correlated. We had a clinical expert perform clinical validity of the learned phenotype relationships.

In the "more common" phenotype set, 82 phenotype pairs were found to have a correlation greater than 0.5 (Figure 4.11). These 82 correlations resulted from 61 unique phenotypes, hence on average each phenotype had more than signi cant correlation with more than one phenotype. Of the 82 reviewed relations 80 (98%) were found clinically valid. One relation rated non clinically valid was the high correlation between a non-disease phenotype for outpatient visits and a non- disease phenotype for inpatient visits. The other non clinically valid relation was between a joint

Figure 4.9: Phenotype coherence scores. Histogram of average phenotype coherence scores assigned by the two clinical expert. Score 1=`not related', 2=`related', 3= `actionable'.

disease phenotype and a phenotype that seemed to be a mix of hepatitis C, liver disease, and obesity.

In the "more rare" phenotype set, 21 phenotype pairs were found be have a correlation greater than 0.5 (Figure 4.12). These 21 correlations results from 23 unique phenotypes. Of the 21 reviewed relations 12 (57%) were found to clinically valid. Most of phenotype pairs that the clinician deemed as unrelated were not very coherent phenotypes which could be expected from phenotypes that were assigned to less of the 5% of the training set.

4.4.2 Hypothesis 2: focus on HIV phenotypes

Of the 250 phenotypes, 73 where identi ed as 'HIV' according to their automatically generated label. The clinical expert evaluation of the these phenotypes showed that most of the identi ed phenotypes represented a routine primary care visit of an HIV patient. Three phenotypes were clear representations of HIV phenotype and two other phenotypes representing AIDS, the development of HIV into a disease. The rest of the phenotypes were of HIV comorbidities (psychiatric, cancer, renal, neurological, etc) mixed with HIV related observations. A few phenotypes captured behavioral phenotypes (substance abuse) and 11 phenotypes were deemed as non-coherent.

Figure 4.10: All signi cant pairwise-positive correlations visualized

4.4.3 Hypothesis 3: focus on non-HIV phenotypes

When categorized into CCS categories according to their ICD label, the learned phenotypes were were found to cover 16 out of the 18 CCS level 1 classi cations (Table 4.2). The two CCS level 1 categories not captured in the phenotype labels pertained to pediatric conditions. Beyond the most prevalent CCS category related to HIV, `Mental Illness' (which include substance use) and `Disease of the circulatory system' were the most frequent disease groups identi ed by the model (Figure 4.13). This nding re ects the high coverage of the learned phenotypes related to

Figure 4.11: Signi cant pairwise-positive correlations evaluated by clinician for clinical correctness.

the types of conditions characteristics of the input population.

4.4.4 Hypothesis 4: types of phenotype- relatedness

Of the 82 relations evaluated, 63 t into the same CCS multi-level classi cation, level 1 category and thus could be inferred using the CCS. However 19 relations were not of the same level 1 category. Out of those 19, 2 were deemed to be unrelated by the clinical expert, 17 relations (21%) were clinically correct and could not be inferred from the CCS and showed more diversity

Figure 4.12: All signi cant pairwise-positive correlations for `rare' phenotypes (de ned as present in less than 5% of the training set).

in the relation type learned: the phenotype for severe HIV and one representing the non-disease ICU visits, as well as comorbidity relations like in the pair for `end-stage renal disease' and `acute respiratory failure.'

4.4.5 Patient-level Content selection

After the model learns 250 phenotypes from the patient population in the HIV clinic, the model can be applied to the data found in a single patient record. Running the model inference on patient

Table 4.2: 250 phenotypes by their CCS category

CCS level 1 category	Number of phenotypes
Infectious and parasitic diseases	83
Mental illness	32
Circulatory system	26
Neoplasms	23
Respiratory system	13
Endocrine; nutritional; and metabolic diseases	12
Digestive system	10
Musculoskeletal system and connective tissue	10
Genitourinary system	10
Nervous system and sense organs	8
Symp; signs; and ill-de ned conditions	8
Blood and blood-forming organs	6
Injury and poisoning	4
Skin and subcutaneous tissue	3
Complications of pregnancy	1
Resid. codes; unclassi ed; all E codes	1
Certain cond. originating in perinatal period	0
Congenital anomalies	0

level data (without re-learning the model parameters) provides a 250 dimensional summarization of the patient record. To summarize the patient record over time, we can run the model inference on the patient data after segmenting the patient data at the desired time granularity. The identi ed phenotype proportions over time is inputted to a sankey visualization presented in Figure 4.4. Each sankey line represents a phenotype identi ed to be relevant in the patient record. The height of each sankey link indicates the proportion of the phenotype in that period. In order to be actionable and avoid information overload the summary showcases the patient's top 5 problems. Top problems are de ned at the phenotypes that are found by the model to have the highest probability among the 250 phenotypes learned by the model. The visualization then illustrates how the proportion of the phenotypes increased, decreased, or stayed the same from one period to the next. As a clinical decision support tool, this summary and visualization of the change in phenotypes identi ed in the patient record could signal to users what health problems the patient possess and how their salience has changed over time.

Figure 4.13: 250 learned phenotypes colored by their labels' corresponding CCS category. Size of the circle indicates proportion of phenotype represented in the training set.

The described approach for patient summarizing using the proposed phenotyping model bene ts from several of the key characteristics of the model. Since the phenotyping model is fully unsupervised the model can easily be utilized for other patient populations by re-training the model on relevant patient data. For instance if patient record summarization was desired for oncology patients, the model can be retrained on oncology patients to learn cancer-speci c phenotypes as well relevant co-morbidity phenotypes. The patient summary bene ts from the high-throughput nature of the model in that the model learns many phenotypes at the same time and is able to summarize the patient record according to all the phenotypes found to be prevalent in the patient record. Finally the model provides a probabilistic summary of the patient record. The generated patient summary is probabilistic in two senses; 1) each data point has a probability of being associated with the phenotypes; and 2) the phenotype assignments to the patient is also probabilistic which can be interpreted as the salience of the associated phenotype in that time period.

To ensure that the model provides a digestible summary of the patient record we analyzed how many phenotypes are required to capture the large majority of the patient record in our training

set. If we nd that the model assigns a large number of phenotypes to each patient record then the proposed summary may still provide too much information to be useful at the bedside. In our analysis we found that more than half of patients in our data set were almost completely described by 1-5 phenotypes (Figure 4.14). The large majority of the remaining patients were described by 6-20 phenotypes. Hence, even though the model is trained to learn a large number of phenotypes (K=250), each patient record is summarized by only a few phenotypes. This indicates that the model has the potential to reduce many thousands of data points in the record of each patient to a list of a handful of problems and how they have changed over time.

Figure 4.14: Number of phenotypes needed to explain 90% of a given patient record. For example, 65% of the patient records in the training set are almost fully explained (90% of data) by 1-5 phenotypes. Each patient record is likely explained by a different 1-5 phenotypes from the 250 phenotypes the model learned from the entire patient cohort.

4.5 Discussion

Evaluation results show the model simultaneously identi es 250 phenotypes with good coherence and coverage. Learned phenotype relatedness were found clinically meaningful and diverse, identifying some relations out of scope of the baseline resource. Our experimentation shows that when training the model on a cohort of HIV patients, the model learns multiple HIV phenotypes that can provide good granularity when used for single-patient problem-oriented summarization. The model was also found to identify a wide range of non-HIV phenotypes, yet commonly encountered in HIV patients. The learned phenotype-phenotype correlations learned from the patient

cohort could be used to group and organize highly-related phenotypes in the patient-level summary, to provide a clearer overview of the patient's problems. In many settings but notably urgent care and emergency settings in particular, patient summarization enabled by this model, could provide clinicians a tool for more rapid understanding of the patient comorbidities, leading to better diagnosis, expedited referrals, and potentially a reduction in over-testing. In Chapter 5 of this dissertation we present the evaluation study of the patient record summarization in assisting clinicians to review patient records more effectively and accurately.

Chapter 5: Usability testing of patient summarization system with target users

Data visualization applications in the health domain have been evaluated using a variety of evaluation methods as reviewed by [302]. These evaluations have been described by two useful dimensions: i) evaluation settings and ii) evaluation measurements. Evaluation settings used in the literature include lab settings with proxies to target users, lab settings with target users, partial roll out or implementation, and full roll-out of system. Previously used evaluation measurements used include unstructured interviews, user surveys (with and without scoring scales), task based measurements, and other outcomes such as health outcomes.

The review recommendations identi ed that evaluation studies going forward should leverage commonly reported metrics for better comparability with existing literature, focus on interaction and work ow, in addition to adopting a phase evaluation strategy. In this Chapter we describe our work on designing and implementing an evaluation study of the patient record summarization system. We sought to follow the recommendations identi ed by [302] while staying consistent with the UCD approach.

In the nal aim of this dissertation we set out to combine the design requirements that we identi ed in Aim 1 and the modeling we developed in Aim 2 to generate a joint phenotype driven summarization system of patient data. After the prototype was complete we ran several heuristic evaluation sessions with two clinical experts. Each expert was asked to explore the summarization tool for 12 HIV patients. Experts reviewed the patient data using the tool as well as the EHR record for each patient. The experts provided feedback about the tool's accuracy, usability, and assisted in designing the next usability study. We then proceeded to conduct a usability study with target users (rather than proxies) to assess the usefulness of the system to support users in the intended

task we aimed to assist in, patient chart review at the bedside. The design of the study tasks in addition to a few summarizer features were informed by the previous heuristic evaluations. The recruited expert for the heuristic evaluation were not recruited for the usability study.

5.1 Prototype no. 2: Combining joint phenotyping and interactive visualization

In this section we describe the tool architecture, the back-end model and front-end interface, as well as the user study of the tool.

5.1.1 Summarization pipeline architecture

The tool consists of three main processes 1) of ine training of a phenotyping model on relevant patient populations; and 2) running the target patient data through the already trained model to obtain a summary; and 3) sending the patient summary from the model to the web-based visualization tool. The rst process is done ahead of time while processes 2 and 3 can be done real time.

5.1.2 Off-line phenotype learning

The model is trained ahead of time to learn phenotype de nitions by leveraging the structures and unstructured data from the records of many patients. For the model to learn phenotypes at the right granularity, the model is trained on a patient population that is similar to the target patient. Each phenotype de nition is composed of diagnoses, medications, laboratory tests, and clinical notes that have been observed to commonly co-occur in the training patient population.

The model also learns the phenotype-phenotype relatedness of each phenotype pair the model produces. By modeling correlations among the phenotypes the model can more accurately re ect the clinical processes of the phenotypes, which are likely to correlate with one another. The learned relationships can also be used on the patient level summary for organizational purposes and for providing higher level abstraction of problems into related problem groupings.

5.1.3 Patient-level phenotype summary

The phenotypes learned from the patient population are then used by the model to assign phenotype proportions to the target patient. To generate a temporal summary of patient phenotype proportions, the model is run on chunks of the patient data at the time granularity the user is interested in. For each time period of interest the model generates a probability distribution over all possible phenotypes. Phenotypes assigned a high probability in a time period indicate that the data types learned to be associated with that phenotype from the population level where present in the patient record during that time period. Since the same phenotype de nitions are used to summarize the patient record over time, the temporal summary of the patient provides how the probability of a single phenotype changes over time. If phenotype probability can be used to proxy phenotype salience, then the temporal summary could pick up on changes in the health status of certain patient problems over time. In addition to the phenotype probabilities over time, the model also provides the information of which data points in the patient record were mapped to what phenotype, allowing for a high level of model transparency, interpretability, and maintained data provenance.

5.1.4 The front-end visualization

The front-end of the summarizer is a web-based interactive visualization built using D3.js, a JavaScript library for producing dynamic, interactive data visualization in a web browser. The visualization is hosted using a simple command-line http server established on a secure server using port-forwarding. This means that the web-application is only viewable from the single machine that launches the server. The front-end visualizer reads in json les that contain the problem proportions the model identi ed for each patient record and time period of interest. The tool interface also reads in the problem assignments of each patient data point, in addition to patient speci c problem labels. For the purpose of the evaluation study all the json summary les were pre-generated but these les can also be generate on the y as more patient data is recorded over time.

The interactive visualization tool is composed of (A) a user selection menu allowing for the selection of the patient id, the number of top problems to view (X), and number of period to view (Y); (B) a summary of the patient problems over time represented in the color-coded sankey diagram; and (C) scrollable boxes showcasing the patient raw data and their assignment to the phenotypes (Figure 5.1). Each colored sankey line represents a patient phenotype. The rectangular sankey nodes for each phenotype represent the phenotype proportion identi ed for the patient during that time period. The tool default is to show the patient's most probable 10 problems in the last 5 time periods. The phenotypes presented in the summary are the X number of phenotypes that had the greatest probability over Y most recent periods selected to be viewed. Hence, if the phenotypes presented in the tool may change a little depending if the user chooses to concentrate on the last 2 years of data versus the last 5 years of data. Even though the phenotype de nitions are learned from the entire population, the tool shows a custom phenotype label for the patient using the most probably diagnosis code assigned to that phenotype that was present in the patient data. If for some reason the patient does not have any diagnosis codes associated with the phenotype, a phenotype is assigned a label using the CCS category from the population level phenotype. Patient data boxes.

The patient raw data summarized are presented in the scrollable boxes on the right. Each of the 4 data type are presented in their own box ordered as follows diagnosis codes, medications, laboratory tests, and words from clinical notes. Each data point in each time period is has a color coded bar behind it indicating the phenotype that the data point was assigned to and the number of times it appeared during that time period. Users can browse all the data points by time period using the drop down menu or one phenotype at a time by clicking on the phenotype sankey diagram.

5.1.5 Data

To train the phenotyping model (described in Chapter 4) we used the EHR data of 6,553 patients from an HIV clinical from NYPH. The patient population was the same one used for the evaluation of the model in Chapter 4 but we now trained the model on the same data conformed

Figure 5.1: Example of patient-speci c summary over ve years. The top 7 most salient problems in between 2014 and 2018 are visualized and how their documentation has evolved through time. The summary is presented at the year level by binning the patient's documentation for that time resolution. The patient has HIV-speci c problems, as well as comorbidities, including asthma, depression, and substance abuse. Relations among the inferred phenotypes are not shown. Dates are changed to maintain patient privacy.

to the OMOP Common Data model [303]. The data spanned 5 years and included the data types: words from clinical notes, laboratory tests ordered, medication orders, and assigned diagnoses codes from across all clinical settings (inpatient, outpatient, emergency). For the purpose of the model training each patient record was restricted to the most recent 4 visits. This was done in order to capture relatively constant period in the patient phenotypes. The nal training data set included the following total data counts and unique vocabulary size in brackets: total words from clinical notes: 37,831,411 (unique: 7,536); total laboratory tests: 1,856,892 (unique: 831); total medications: 144,975 (unique: 1,970); and total diagnosis codes: 172,550 (unique: 1,991). For patient summarization, patient data spanned the ve most recent years of data. Patient data was aggregated at the yearly level by enumerating the number of times each observation appeared in the patient record during the year.

5.2 Evaluation methods

5.2.1 Usability Study with HIV clinicians

The goal of the evaluation study was to assess the ability of the tool to support clinicians in reviewing patient charts. Subjects were recruited via an invitation email sent to clinicians that regularly care for HIV patient from NYPH's Division of Infectious Diseases. Inclusion criteria for participants was that they were practicing clinicians at NYPH in east or west campus and that regularly care for patients with HIV. Participation was voluntary and compensated. All subjects were experienced users of the EHR system at NYPH but had never seen the summarization tool before.

Patient case selection

Cases for patient review were selected from the NYPH HIV patient clinic and selection focused on patients with a long history of care with multiple comorbidities, and a combination of outpatient, inpatient, and ED visits. Patients data included in the study ranged between 4 to 5 years of clinical follow-up at NYPH, although patients had longer histories that were not included in the evaluation. Over the included period all of the patients had seen several medical providers, numerous visits, and signi cant documentation with a sizeable number of diverse problems. All records were checked to ensure that the physician participant reviewing the patient case had not cared for the patient in the last two years.

Study protocol

Each participating clinician was randomly assigned to review 4 patient cases, with 10 minutes dedicated to each patient case to simulate realistic time constraints. Clinicians reviewed 2 patient cases with the baseline system which is the EHR at NYPH (Condition A) and 2 patient cases with the aid of the summarizer (Condition B). The order in which participants reviewed each patient and under what condition were alternated (Table 5.2).

Table 5.1: Summary statistics on electronic health data of the 8 study patient cases

	Mean	Median	Range
Number of visits at NYPH	88	66	47 -176
Number of notes	764	487	128- 2,112
Total laboratory tests	2,233	1,584	364-4,906
Unique laboratory tests	475	425	288-760
Total diagnosis codes	1,756	1,911	1,177- 1,975
Unique diagnosis codes	306	227	63- 751
Total medications	1,716	1,778	1,461- 1,887
Unique medications	186	113	21- 539
Words from clinical notes	22,614	15,839	5,323-51,347
Unique words	7,526	7,529	7,503 - 7,535

Table 5.2: Study protocol of 2 groups of 8 clinicians (total of 16) reviewed 4 patient cases each (total of 8 patient cases). Bold patient ids indicates the study condition with summarizer (Condition B) and the non-bold represents the baseline use of the EHR (Condition A).

Clinician	Case 1	Case 2	Case 3	Case 4
Clinician 1	Pt F	Pt H	Pt D	Pt A
Clinician 2	Pt F	Pt H	Pt D	Pt A
Clinician 3	Pt H	Pt A	Pt F	Pt D
Clinician 4	Pt H	Pt A	Pt F	Pt D
Clinician 5	Pt D	Pt F	Pt A	Pt H
Clinician 6	Pt D	Pt F	Pt A	Pt H
Clinician 7	Pt A	Pt D	Pt H	Pt F
Clinician 8	Pt A	Pt D	Pt H	Pt F
Clinician 9	Pt G	Pt E	Pt C	Pt B
Clinician 10	Pt G	Pt E	Pt C	Pt B
Clinician 11	Pt E	Pt B	Pt G	Pt C
Clinician 12	Pt E	Pt B	Pt G	Pt C
Clinician 13	Pt B	Pt C	Pt E	Pt G
Clinician 14	Pt B	Pt C	Pt E	Pt G
Clinician 15	Pt C	Pt G	Pt B	Pt E
Clinician 16	Pt C	Pt G	Pt B	Pt E

Table 5.3: Participant Tasks under Condition A (EHR) and Condition B (Summarizer)

	Condition A	Condition B
Task 1.1	Generate problem list using EHR alone (PL-EHR)	Generate problem list using summarizer tool alone (PL-Viz)
Task 1.2	N/A	Review patient data with EHR, revise PL-Viz from task 1.1 if needed (PL-Viz+EHR)
Task 2	Answer two questions about patient using EHR alone (Q-EHR)	Answer two questions about patient using tool alone (Q-Viz)
Task 3	Generate 1-2 sentence patient summary with the EHR alone, noting if each problem is improving, stable, worsening when possible (Summary-EHR)	Generate 1-2 sentence patient summary using the tool alone, noting if each problem is improving, stable, worsening when possible (summary-Viz)

Evaluation tasks

For each patient case the clinicians were asked to perform 3 main tasks (Table 5.3). Task 1.1 was to generate a patient problem list using the EHR alone if under Condition A (referred to as PL-EHR) and using the summarizer if under Condition B (PL-Viz). Since the summarizer is not intended to replace the EHR but to aid clinician to gain an overview of the patient, participants were asked to con rm the problem list they generated under condition A and to revise it if you needed to (PL-Viz+EHR). Task 2 was to answer two patient speci c questions regarding the patient's treatment or medical history (e.g. Q1: In 2013 patient sought physical therapy, what was it for? and Q2: What was the HIV medication regimen in 2014?). Question answers under Condition A are referred to as Q-EHR and under Condition B as Q-Viz. Participants did not verify their answer to the questions with the EHR when under Condition B. Finally, in Task 3 participants were asked to generate a 1-2 sentence patient summary of the patient case. In the summary participants were asked to note the status of mentioned problems, that is if problems were improving, worsening, or stable. If they were unable to note the status of a problem given using the EHR (Condition A) or the summarizer (Condition B) they could note this in their answer. After reviewing 4 patient cases participants were asked to complete a usability questionnaire following the System Usability Scale (SUS) [304] and provide free-form feedback regarding the tool.

Table 5.4: SUS Questionnaire

Question
I think that I would like to use this system frequently
I found the system unnecessarily complex.
I thought the system was easy to use.
I think that I would need the support of a technical person to be able to use this system
I found the various functions in this system were well integrated
I thought there was too much inconsistency in this system
I would imagine that most people would learn to use this system very quickly
I found the system very cumbersome to use
I felt very con dent using the system
I needed to learn a lot of things before I could get going with this system

Evaluation metrics

To assess performance under each study conditions data measurements collected included: a) -to-completion of Tasks 1.1 through 3, b) problem list precision and recall for problem lists PL-EHR, PL-Viz, and PL-Viz+EHR. Precision and recall were calculated in comparison to gold-standard lists generated by two clinicians; c) correctness of clinical questions for Q-EHR and Q-Viz, scored in comparison to gold-standard solutions generated by two clinicians. Each question was scored out of 100%. Questions were scored with a score of zero if were incorrect or were not answered, a score of 50% if partially correct, and 100% if perfectly correct; d) patient summary score, ranging between score=0 if summary was not completed, score=50 if completed but did not note status, and score=100 if successfully noted problem status; e) usability of the summarization is assessed through a post-study questionnaire adapted from the System Usability Scale (SUS) which is a standard tool to measure usability [304].

The 10 items asked in SUS, users are asked to score one of ve responses Strongly Agree (score=5), Agree (score=4), Neutral (score=3), Disagree (score=2), and Strongly Disagree (score=1).

Data collection This study used Open Broadcaster Software (OBS) which allows to record multiple screens and audio. All participants used the same computer connected to two large external screens and had access to the EHR on one screen and the summarizer interface on the other. All answers were entered by the participants in a Word document.

Statistical analysis

Two clinicians completed an answer key for the medical questions and constructed gold-standard problem lists for each patient case. Question responses and problem lists generated by the study participants were compared to the gold-standard solutions and problem list accuracy and questions correctness were calculated. Problem lists accuracy metrics calculated included precision and recall. Statistical signi cance between the recall and precision of the two study conditions is tested using the non-Parametric test Wilcoxon signed-rank test. statistically signi cant difference between the conditions for the question correctness was calculated using a Wilcoxon signed-rank test as well. A comparison of the average to completion under the treatment and the control conditions is evaluated using Multi-source Analysis of Variance (ANOVA).

5.3 Results

Sixteen clinicians participated in the evaluation study, 14 infectious disease clinicians from the NYPH HIV clinic, 1 pediatrics infectious disease clinician, and 1 hospital internist from Weill Cornell. In total the participants included 5 Assistant Attendings, 2 Associate Attendings, 4 Nurse Practitioners, and 5 Clinical Fellows. Subjects completed 4 patient cases for a total 64 case reviews, 32 with the aid of the summarizer tool and 32 with the EHR alone. Since the 16 clinicians were split into two groups, to review 4 different patient cases each patient case was reviewed by 8 clinicians, 4 times with the aid of the summarization tool and 4 times with the EHR. The time constraint set to 10 minutes per patient case was found to be quite strict and the participants needed to complete the problem list, clinical questions, and summary in a brisk pace. Out of the 64 case reviews, 7 cases could not be completed in full due to time running out. In each of the 7 cases where the participants could not nish all sub-tasks in time only as single sub-task was missed (6 patient summaries and 1 clinical question; 4 under the EHR condition and 3 under the summarization condition). For the result analysis, a time penalty of 10 minutes was assigned to the sub-task when participants could not get to it in time in order to avoid recording 0 minutes for the tasks which would have biased

the results. The solution to the sub-task is recorded as wrong.

5.3.1 Task 1: Problem lists

Problem list construction using Condition A (EHR) versus Condition B (Summarizer)

To construct the problem lists under Condition A (EHR) that most used strategy was to looked for the last note in the time period of interest that was written by the patient's primary care physician (PCP). Once they found that note they skimmed through it, noting any mentioned problems. Once constructing the list of problems, participants often looked for medications mentioned in the same note to verify they were consistent with the listed problems. At times, when there was a mentioned problem but with no relevant medication they removed the problem from the list. Some participants scrolled through the patient notes to con rm the identi ed problems with another note. One or two participants also searched for notes written by specialists, stating that they would likely have problem lists with a different focus and thus were interested to look at those for completeness. Since participant recognized the name of the PCP of these patients they often search for the clinicians notes by name. They also often mentioned that they know that the speci c PCP makes detailed notes and thus they don't need to view any other data source in order to construct the patient problem list.

Under Condition B (summary system) participants had two stages for the completion of the problem list: Task 1.1 entailed generating a list with the summarizer alone; and Task 1.2 was to look at the patient EHR to edit the problem list from Task 1.1 if they wanted to. Strategies taken to construct PL-Viz varied somewhat between participants. Three main user behaviours were identi ed. Strategy 1 included just reading off the problem labels off of the sankey diagram, without investigating the underlying data assigned to each problem. Often they choose to ignore problems for which the sankey link was very thin in last time period presented, indicating that no patient data was mapped to the problem in that year. In strategy 2 participants took a very different approach and clicked on every problem link and on each time period, looking through the patient data assigned to each problem quite a bit. In strategy 3, users only glanced at the sankey diagram

but directly navigated to the option of viewing all of patient data by period.

Problem list accuracy using Condition A (EHR) versus Condition B (Summarizer)

Problem list accuracy was calculated in comparison to an expert-generated gold-standard problem lists created by two clinicians (Figure 5.2). Problem list precision was found to be fairly similar between the study conditions and differences were not found to be statistically signi cant. The average precision for PL-Viz was the highest with precision of 0.66 (median precision=0.667). Precision was slightly lower for PL-EHR at 0.65 (median precision=0.667). Finally, the lowest precision was obtained for PL-Viz+EHR at 0.61 (median precision=0.62). That is the average (and median) problem list precision under Condition B was reduced in Task 1.2 from Task 1.1, when clinicians were asked to review the patient EHR and make any edits to the patient problem they generated with the summary alone (PL-Viz). Problem list edits made in Task 1.2 were largely including more problems to the list. Only one clinician removed a problem previously identi- ed in PL-Viz. The additional problems added to the list in PL-Viz+EHR were often not on the gold-standard list which increased the false positive rate, reducing precision.

By contrast to precision, PL-Viz+EHR had the highest average recall at 0.9 (median recall=1). PL-EHR had the second highest average recall of 0.84 (median recall=0.8), followed by the average recall of 0.75 (median recall=0.77) for PL-Viz. Hence the problem lists PL-Viz+EHR were found to be most complete, with the fewest false negatives (hence missed problems). When performing non-parametric test for the differences in the median recalls they were found not to be statistically signi cant.

5.3.2 Task 2: Clinical questions

Question answering using Condition A (EHR) versus Condition B (Summarizer)

Most of the participants attempted to answer the clinical questions by searching for clinical notes that may have this information. This general strategy was used even when the question was regarding medication that was used certain problems or about procedure performed. Some

Figure 5.2: Boxplot gure of problem list precision and recall by condition.

questions were easier to nd in the notes, such as the HIV medication regimen the patient was on, since most of the participants knew that this information was well documented in the clinical notes of the patient. However when the question were regarding medication treatment the patient received while in an inpatient setting (such as an IV drip) this was harder or even impossible to nd in the notes. In which case users scrolled through the medication orders in the EHR. Questions that had any temporal aspect to it, such as assessing the continuity of medication was also dif cult to do under Condition A. To obtain this information users often selected two notes to view that were a few years apart in an attempt to identify any change in the treatment of the patient.

Under Condition B questions were much easier to answer, even for users that were less secure in their use a of the system. Participants often navigated to the problem that was most relevant to the question in the sankey diagram and browsed its supporting data, either for the entire time period presented or speci cally to the year of interest if speci ed in the question. At times when the information was not found under the expected problem the users navigated to option of viewing all of the patient data for each time period and browsed the data in this manner to nd the information they needed.

Question-answer correctness using Condition A (EHR) versus Condition B (Summarizer)

Most of the clinical questions were answered correctly under both condition, with the median questions score being 100 under both conditions A and B. The average question accuracy score was higher for Q-Viz at 91 (standard deviation=17.6) than 85 (standard deviation=21) for Q-EHR. Accuracy differences were not found to be statistically signi cant. In 3 instances (out of 64 question responses) clinicians were not able to nd the answer to a patient question under Condition A. This did not occur under Condition B, but one participant did run out of time while completing one of the questions.

Figure 5.3: Boxplot gure of clinical question answer scores by study condition. The mean question score by condition is showcased by the red dot and number label.

Task 3: Patient summary

Summary contraction using Condition A (EHR) versus Condition B (Summarizer)

Under Condition A, participants often searched for the latest note written by the patient's PCP and searched inside the note for the case summary. Participants often wrote verbatim what the note stated. If the patient summary on the notes indicated any status change they did so as well in their response to the evaluation Task. When the status of the problem was not mentioned they failed to specify the status as well. Several participants navigated to view the laboratory results of the patients to note the status of their viral load to indicate how controlled their HIV was.

Under Condition B participants used the sankey diagram to indicate any notable problems. To indicate the status of the problem, participants tried to note if there were any medication changes to indicate a change in the problem status. No participant blindly trusted the width of the sankey diagram as an indication of the change in the problem status, especially for chronic problems. Some acute problems were noted if the sankey diagram had a sudden increase in its width and the underlying data supported the spike.

Summary quality using Condition A (EHR) versus Condition B (Summarizer)

The average summary quality under Condition A was slightly higher than under Condition B, with average quality score of 62.5 compared to average score of 61 (respectively). The lowest score under both conditions was 0, while under Condition B the lower 25th percentile score was 37.5 while under Condition A it was still 0. However, the median quality score under Condition A was 100, and only 50 under Condition B. Four clinicians ran out of time and were not able to complete patient summaries for one of their patient cases, 2 instances were under Condition A and 2 were under Condition B.

Figure 5.4: Boxplot gure of patient summary scores by study condition. The mean summary score by condition is showcased by the red dot and number label.

Time-to-completion

Total time-to-completion of the problem lists under Condition B (PL-Viz+EHR) was longer than under Condition A (at 4.58 and 3.2 minutes on average). Since under Condition B users took time to construct the problem list using the summarizer and then looked into the EHR to verify and edit if they wished the slightly longer time to completion was expected. By contrast, question answering was statistically signi cant faster under Condition B than under Condition A (at 1.8 and 3.07 minutes on average). The patient summary took an average of 2.7 minutes under Condition A and 2 minutes under Condition B. According to the ANOVA analysis of the time to completion of the tasks, factors found to statically affect the time to completion were the patient, the clinicians, the task, the study condition interaction with the task. These results make sense as some patient cases were more dif cult than others, variation between clinician was noticeable,

and the evaluation condition in uenced the time-to-completion and the directionality of that effect depended on the task.

Figure 5.5: Time to completion of each task by condition. The three outliers (1 for Q-Viz, 1 for Summary-EHR, and 1 for Summary-Viz) at 10 and 12 minutes are users that did not complete those tasks in time. In those cases, the time to completion was changed to 10 minutes of for the task as a penalty.

Usability scores

The usability survey of the system found that half of the participant (n=8) indicated they would use the system frequently, 5 were unsure, and 3 indicated they would not. The system was largely found to be not complex, easy to use, consistent, and with well integrated features.

Participant feedback

Participant feedback was analyzed and the following themes were identi ed:

Figure 5.6: Usability survey results. Question scores were normalized so that low scores (1-2) express negative sentiment towards the usability of the summary, high scores (4-5) indicate favorable sentiment towards the summary, and a score of 3 is neutral.

Interface Design and temporal awareness. Several clinicians (C6, C7, C8, C12, C13, C14) emphasized the utility of such a quick temporal snapshot of patient data, which makes it easier to navigate one period at a time. One clinician noted they liked the problem list view of the sankey diagram, with the detailed data on the side (005). However, another clinician noted that the sankey diagram could take less space in the visualization (001). One clinician (C12) noted the interface was "helpful because it displays a lot of data very quickly and succinctly that is difficult to get out the chart". Several clinicians (C5, C8) noted they liked the color coding of the problems and problem thickness, and others (C5, C7, C11) mentioned they liked red font for new data points . Visualization novelty and time constraints. One clinician (C5) emphasized they were not used to reviewing visualizations for patient data other than for laboratory results. One clinician (C1) noted that because of the time constraint they didn't get a chance to explore the entire summary in detail. A few participants (C6, C7, C13) stated that they needed more time with the tool to become familiar with it in order to gain an understanding of what it gets right and wrong. Another (C16) stated that they need more time to investigate how the sankey line width correlated with the patient state.

Input data, additional features, and problematic features. Several clinicians (C5, C8) noted they liked that the summarization used multiple input types, which made the summary more robust.

Clinicians emphasized the importance of laboratory tests, medications, and clinical notes but were skeptical in the utility of the diagnosis codes which often do not accurately re ect the patient state. The laboratory tests would need to be linked to numerical results, the medication names linked to dosages, and the clinical words need to be ltered in order to be useful (C1, C6, C2, C8). One clinician (C13) noted they wanted more transparency of where the data was coming from and another (C1) noted they wanted to know how many visits were represented in each time slice. Other features that were requested by a few (C15, C11) included sorting capabilities of the patient data and search functionality. One clinician (C7) also asked to include procedure data, and to split the data to medical and surgical. Another clinician (C3) wanted a separation of chronic and acute problems. One clinician noted that the summarizer may under-represent problems that are poorly documented in clinician documentation such as substance abuse.

Problematic features in the visuals included the sankey links going up and down. Some clinicians (C5) said they expected the ups and downs to have meaning, even though they did not have one. This also made it dif cult for some users to follow the contours of the line making them click on the wrong problem and made it dif cult read the problem labels. The rst period of the data was hard for some clinicians (C5, C11) to navigate to, in addition to understanding when they were viewing problem speci c data and when they were viewing all the data.

Trade-off between transparency and perceived accuracy. The general sentiment of participants was that the problem-level abstraction of the patient data were mostly consistent, with the main problems identi ed for the patient being correct (C3, C4, C16, C5). However, several participants (C1, C3, C6, C14, C10, C11) agreed that the exposed modeling inference for each patient data point had some low-level inconsistencies, mapping certain data points to unrelated problems (e.g. hypertension medication assigned to the HIV). Consequently, some clinicians (C3, C16) noted that even few inconsistencies would make them lose trust in the system, causing them to spend more time looking through the raw data or even conclude that they would not use the system as is (C6). Other clinicians (C4, C2) believed the system was consistent or that they were used to dealing with EHR inconsistencies and `messiness'. One clinician (C8) noted that there was

enough information provided in the tool that allowed them to easily apply clinical judgment and decide if presented information was trustworthy. The same clinician also noted that they would be interested in a system that would improve and learn from user corrections, through a human-in-the loop approach, with the assistance of an easy-to-use interface.

A few clinicians (C9, C16) responded they were more comfortable searching for information in the EHR since they were more familiar with it. However, the summary tool would help if they were not familiar with the EHR. They added that they have a good sense of what information they can trust and cannot trust in the EHR system. Two clinician (C3, C9) expressed that he would worry about the tool not being comprehensive and may be missing data. Others (C3, C16) noted they would want to verify the summary information with the EHR. One clinician (C13) noted that the summarizer would be a useful addition for any EHR but that clinicians would need to build it into their work ows.

Disease status, change overtime, prioritization, and context. According to several participants (C2, C7, C8, C10, C16) the width of the sankey links were noted to help bring attention to new problems, events, or are ups of a problem. One clinician (C8) noted "I really love how you can see over time how things change in terms of what is receiving attention and what are the predominant themes in the patient record." Multiple participants noted (C1, C7, C10) that this type of information would be dif cult to manually identify in the patient EHR or using other data representations such as lists. Participants (C2, C10) added the tool was useful in comparing problem dominance and save time on assessing prioritization. Others (C2, C12) said the summarizer did a really good job at weeding out noisy data and helped identify relevant information. However, participant (C7) wished they could shuf e the rank of the problem by prevalence in a certain time period and two clinicians (C14, C15) wanted resolved problems to be removed from the visual.

However, several clinicians (C3, C6, C7) noted that in order to make judgments regarding the status change of a problem they would need additional context provided in relevant clinical notes. Although it was acknowledged that currently it could be hard to nd relevant notes using the EHR. Others (C4, C6, C14, C7) pointed out the need for laboratory results or medication changes.

5.4 Discussion

The study results identify that fully unsupervised probabilistic phenotyping coupled with interactive visualization can generate clinically meaningful patient summaries for complex patients such as those with HIV. The summaries were robust to different patients as they identi ed the HIV condition in each patient but also additional diverse comorbidities in each patient case. The patient main problems for each patient were found to be largely consistent with the gold-standard problem lists. The visual representation of problems using the color coded sankey diagram that widened and narrowed over time was interpretable and assisted clinicians identify salient patient problems over time. The heavy use by participants of the sankey diagrams, the detailed lists of the patient data, and transitioning from viewing all of the patient data to ltering to problem related data showcased the utility of allowing for different levels of detail granularity and focus in the summary.

Exposing the model inference for each data point in the patient record to the user through the summarizer interface provided useful evidence regarding the accuracy and robustness of each problem identi ed by the summarizer. For instance problems that had only a few data points mapped to them or mostly unrelated data points helped clinicians identify problems that could be ignored in the summary. Although exposing such low-level inference to the user helped with interpretability and explainability, it also exposed some low-level modeling errors for individual data points. Inaccuracies in data assignments in uenced the participants' perception of the summary consistency to differing degrees which points to an interesting trade off between transparency and perceived accuracy of machine learning based CDS. Some clinicians noted using clinical judgment with the information provided by the summary they could use the summary con dently, the same was they have to do with the EHR which also contains inaccuracies. Other clinicians noted that even slight inaccuracies in the data assignments made them lose their trust and con dence in the system.

Although there were differences between the problem list and question accuracy between Condition A and B, there were not found to be statistically signi cant. This is likely a limitation of the sample size, which is a common limitation of evaluation studies that are dif cult to run in

large scales [12]. The study showed that the average precision of the problem lists under all study conditions were in the 0.61-0.66 range, which is not very high. Moreover, precision was found to decrease when users revised the problems lists they found with the tool (PL-Viz) once looking at the EHR (PL-Viz+EHR). This was largely due to the addition of patient problems that were not in the gold-standard problem lists. The extra problems that were not in the gold-standard were not necessary wrong but were not thought to be signi cant enough to be put in the gold-standard. This highlights a few things i) constructing patient problem list is a taxing task, especially for complex patients that have a lot of documentation in the EHR. This may have caused participants to generate very long problem lists, basically adding any problem that may be signi cant for the sake of completeness. When generating the gold-standard lists the clinical experts were not under time constraints and had more time to lter the lists to only the problems they thought were signi cant. Furthermore, the low precision rates also point to the general subjectivity of problem lists and the dif culty of assessing their accuracy in comparison to a pre-generated gold-standard. To avoid this limitation it may be better to restrict problem lists to the top 5 or 10 problems in the patient case instead of assigning no limitation on the number of noted problems. Moreover asking the participants to rank the problems in order of importance could allow to assess precision@K and recall@K, metrics that are commonly used in information retrieval tasks in which the order of the elements matter.

The gain in ef ciency in answering the clinical questions using the summarizer versus the EHR highlights the dif culty of identifying certain types of information in the EHR. The summarizer was able to concentrate the multi-source information in one place, making it easy to identify medication by year, and identify notable events by sudden width change in the temporal sankey diagram. However it was noted from the participant feedback that still a lot of the patient story, data context, and assessment of the treating clinicians was missing form the tool and thus required access to the EHR. Hence it was agreed upon by several participants that the summary could provide a great overview and navigational tool for the raw EHR. Greater integration of the summarization system with the EHR, allowing for seamless navigation from the summarizer components into the patient

notes and laboratory results could further improve the utility of such a tool in both accuracy and ef ciency when reviewing patient historical data.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

Complex patient are becoming more and more of a challenge to the health care system given the amount of care they require and the amount of documentation needed to keep track of their state of health and treatment. Record keeping using the EHR makes this easier but mounting amounts of patient data also means that clinicians are faced with information overload. Information overload has been shown to have deleterious effects on care, with increased safety concerns due to missed information. Patient record summarization has been a promising mitigator for information overload. Subsequently, a lot of research has been dedicated to record summarization since the introduction of EHRs. In this dissertation we examine whether unsupervised inference methods can derive patient problem-oriented summaries, that are robust to different patients. By grounding our experiments with HIV patients we leverage the data of a group of patients that are similar in that they share one common disease (HIV) but also exhibit complex histories of diverse comorbidities. Using a user-centered, iterative design process, we design an interactive, longitudinal patient record summarization tool, that leverages automated inferences about the patient's problems. We nd that unsupervised, joint learning of problems using correlated topic models, adapted to handle the multiple data types (structured and unstructured) of the EHR, is successful in identifying the salient problems of complex patients. Utilizing interactive visualization that exposes inference results to users enables them to make sense of a patient's problems over time and to answer questions about a patient more accurately and faster than using the EHR alone.

6.2 Contributions

At the point of care clinicians have a lot of data at their disposal but in order for them to utilizes this information to provide better care for their patients they needs assistance from CDS. To enable scalable and generalizable CDS, a new generation of systems need to leverage advancements in machine learning. However, for the systems to be useful for clinicians these systems need to be interpretable and user friendly, requiring the bene ts visualization. This dissertation contributes by identifying gaps and opportunities for machine learning and visualization in CDS. It also investigates the use of such an approach for patient record summarization of complex patients. The contributions of the work in more detail are the following:

- Identi ed gaps and opportunities for the use of ML and dataVis in CDS. Through an expansive review of CDS literature we classify works into three CDS types (Infobutton CDS, CSO CDS, and Alert CDS) and by the methods they utilized (Heuristics, ML, and visualization). We identify gaps and opportunities for the use of ML and dataVis in all three types of CDS. Speci cally, highlighting the need for greater utilization of these methods for CSO CDS including patient records summarization.

- Generated design requirements for longitudinal summarization of patient records. Through an iterative user-centered design approach we collected the information needs from longitudinal patient summarization. We leveraged previous literature, clinician interviews, and early usability testing with clinicians of an initial prototype. We translated the identi ed information needs into a set of design requirements. The design requirements translated to a set of visual design decision, modeling requirements, and interactivity features.

- Developed and evaluated an unsupervised probabilistic model to jointly learn phenotypes and phenotype relationships using multi-source patient data found in the EHR. We show that the model when trained on the EHR data of many HIV patients is able to learn clinically meaningful phenotype and phenotype relationships. The model was able to learn

multiple HIV phenotypes as expected when training on a patient population of HIV patients and also learn many non-HIV phenotypes. When comparing to knowledge-based baseline of disease hierarchies the model was able to learn phenotype relationships that were clinically meaningful but that could not be inferred from the baseline.

- Expanded Laplace variational inference to accommodate multi-dimensional data in non-conjugate Bayesian modelsPreviously proposed Laplace variational inference, is a generalized form of variational inference that can handle non-conjugate models. In this dissertation we generalize the proposed Laplace variational inference even further to allow for multiple input types. This makes the model inference especially relevant to clinical data which contains many different data types that are important in combination when seeking to learn robust models on clinical data.

- Demonstrated the potential of reproducible, scalable, interpretable approach for patient record summarization through the use of unsupervised join learning of phenotypes and interactive data visualizationsIn the Aim 3 of this dissertation we evaluate the usability of summarization system that leverages a fully unsupervised computational method for patient record summarization, coupled with interactive visualization. The approach was found to successfully support the task of patient chart review of patient with complex medical histories and multitude of chronic and acute comorbidities. The system in combination with the EHR was found to support clinicians in generated problems lists with high recall rates and more accurate question- answering. Through participant feedback we identi ed next iteration on the design requirements of the summarization system.

6.3 Limitations

We acknowledge that this dissertation posses from several limitations. The limitations include the following:

- Generalizability. The aim of the research is to assess the viability of using unsupervised

machine learning and visualization to assist in the summarization of records of patients that suffer from chronic disease and multimorbidty. To do so we elected to focus our experimentation on patients with HIV, as an extreme example of complex patients with multimorbidty. The generalizability of our ndings to other patient populations and other clinician specialization was not assessed in this dissertation.

- Computational Baseline. The performance of our developed computational phenotyping model used for patient problem inference was not compared to other computational methods. The focus of this dissertation is to assess a computational method that satis es all the desirable criteria of patient summarization as listed in the design requirements discussed in Chapter 3. No other computational method fully satisfy those criteria and thus we elected to evaluate our method using expert scoring and comparison to expert generated groupings of disease (the CCS). Other computational methods that have been evaluated for unsupervised phenotyping have failed to utilize both structured and unstructured data or do not explicitly allow for phenotype interrelatedness, which we thought important when summarizing patients with multimorbidity.

- Visualization alternatives. The effectiveness of the sankey visual representation of patient problems and their change over time was not compared to other possible visual representations of the same information. While many other works have experimented with the representation of temporal summaries of patients. No proposed visualization was set to explicitly expressed change in salience of multiple problems over time.

- Input data selection and processingPatient summarization was performed using multiple input types including diagnosis codes, laboratory test names, medication names, and words from clinical notes. The selection of different input types and the their processing could have in uenced our ndings. For example words from clinical notes were processed one word at a time and thus lost some of their context such as negation. This may have some deleterious effect on our ndings.

6.4 Future Work

This work showcases the great potential of our developed approach for patient record summarization available at the point of care. However, as the mentioned limitations indicate there are several interesting directions of future work. Those include the following:

- Summarize other patient types Future work should evaluate the effectiveness of the proposed patient summarization approach for different patient populations and different clinical specializations. Interesting application of the method would be to other patient populations that suffer from chronic disease and multimorbidity such as cancer patients, diabetes patients, among others. It would also be of interest to assess the effectiveness of this method on general population patients that may be less complex but may be more heterogeneous in their problems.

- Include additional data types. The current phenotyping model and patient summarization was performed using diagnosis codes, clinical words, laboratory test order, and medication orders. However, the model and thus the summarization can run on many more data sources such as procedure codes or laboratory value ranges. Although previous experimentation to expand found that including other variables did not improve the performance of the model in terms of automated metrics for model fit such as held-out log-likelihood [93], adding additional data types to the patient summary through the model could be especially useful for tool users. Adding additional data types was noted by a few participants of the evaluation study as being useful additions to the tool.

- Further iterate on summary design. The final usability study presented in this dissertation identified usability weaknesses and more desired features that could improve the utility clinician get from the summarizer. Features include integrating the tool with the EHR to provide a seamless navigation between the summary and the patient data.

- Formally evaluating sankey width and patient state correlation.More formal evaluation

should be conducted to assess how well phenotype salience, as inferred from the probabilistic phenotyping model, correlates with testable bio-markers of disease state. Examples include viral load for HIV, and other laboratory tests for renal disease, cardiac disease, and diabetes.

References

[1] CDC. About Chronic Diseases. 2019. (Visited on 01/19/2020).

[2] R Navickas, VK Petric, AB Feigl & M Seychell. Multimorbidity: What do we know? What should we do? Journal of Comorbidity 6.1 (2016), 4–11.

[3] K Barnett et al. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet (London, England) 380.9836 (2012), 37–43.

[4] CDC. Health and Economic Costs of Chronic Disease. 2019. (Visited on 03/19/2020).

[5] O Farri, DS Pieckiewicz & AS Rahman. A qualitative analysis of EHR clinical document synthesis by clinicians. AMIA Annual Symposium Proc. American Medical Informatics Association, 2012, 1211–1220.

[6] L Samal, A Wright, BT Wong, JA Linder & DW Bates. Leveraging electronic health records to support chronic disease management: the need for temporal data views. Informatics in Primary Care 19.2 (2011), 65–74.

[7] LL Weed. Medical Records That Guide and Teach. New England Journal of Medicine 278.11 (1968), 593–600.

[8] C Plaisant, B Milash, A Rose, S Widoff & B Shneiderman. LifeLines: visualizing personal histories. Human factors in computing systems common ground. New York, New York, USA: ACM Press, 1996, 221–ff.

[9] SM Chowdhry, RG Mishuris & D Mann. Problem-oriented charting: A review. International journal of medical informatics 103 (2017), 95–102.

[10] J Gallant, PY Hsue, S Shreay & N Meyer. Comorbidities Among US Patients With Prevalent HIV Infection—A Trend Analysis. The Journal of Infectious Diseases 216.12 (2017), 1525–1533.

[11] SG Deeks. HIV Infection, In ammation, Immunosenescence, and Aging. Annual review of medicine 62 (2011), 141–155.

[12] M Altman, TT Huang & JY Breland. Design Thinking in Health Care. Preventing Chronic Disease 15 (2018).

[13] A for Healthcare Research & Quality. Learning Health Systems. (Visited on 05/15/2018).

[14] RA Greenes et al. Clinical decision support models and frameworks: Seeking to address research issues underlying implementation successes and failures. Journal of Biomedical Informatics 78 (2018), 134–143.

[15] A Rajkomar, J Dean & I Kohane. Machine Learning in Medicine. New England Journal of Medicine 380.14 (2019), 1347–1358.

[16] DA Norman & SW Draper. User Centered System Design; New Perspectives on Human-Computer Interaction. USA: L. Erlbaum Associates Inc., 1986.

[17] G Ramos et al. Emerging Perspectives in Human-Centered Machine Learning. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–8.

[18] JA Fails & DR Olsen. Interactive machine learning. Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03. Miami, Florida, USA: Association for Computing Machinery, 2003, 39–45.

[19] T Kulesza, M Burnett, WK Wong & S Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15. Atlanta, Georgia, USA: Association for Computing Machinery, 2015, 126–137.

[20] PY Simard et al. Machine Teaching: A New Paradigm for Building Machine Learning Systems. arXiv:1707.06742 [cs, stat] (2017).

[21] F Doshi-Velez et al. Accountability of AI Under the Law: The Role of Explanation. SSRN Scholarly Paper ID 3064761. Rochester, NY: Social Science Research Network, 2017.

[22] R Guidotti et al. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys 51.5 (2018), 93:1–93:42.

[23] B Goodman & S Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine 38.3 (2017), 50–57.

[24] C O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. USA: Crown Publishing Group, 2016.

[25] RL Wears & M Berg. Computer Technology and Clinical Work: still waiting for Godot. JAMA 293.10 (2005), 1261.

[26] R Caruana et al. Intelligible Models for HealthCare. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2015, 1721–1730.

[27] A Rajkomar, M Hardt, MD Howell, G Corrado & MH Chin. Ensuring Fairness in Machine Learning to Advance Health Equity. Annals of Internal Medicine 169.12 (2018), 866–872.

[28] MA Musen, B Middleton & RA Greenes. Clinical Decision-Support Systems. Biomedical Informatics London: Springer London, 2014, pp. 643–674.

[29] WW Stead & WE Hammond. Computer-based medical records: the centerpiece of TMR. M.D. computing : computers in medical practice 5.5 (1988), 48–62.

[30] SM Powsner, CA Riely, KW Barwick, JS Morrow & PL Miller. Automated bibliographic retrieval based on current topics in hepatology: hepatocrit. Computers and biomedical research 22.6 (1989), 552–64.

[31] SM Powsner & PL Miller. Automated online transition from the medical record to the psychiatric literature. Methods of information in medicine 31.3 (1992), 169–74.

[32] RA Miller, FM Gieszczykiewicz, JK Vries & GF Cooper. CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources. Symposium on Computer Applications in Medical Care Proc. 1992, 86–90.

[33] JJ Cimino, SB Johnson, A Aguirre, N Roderer & PD Clayton. The MEDLINE Button. Symposium on Computer Applications in Medical Care Proc. American Medical Informatics Association, 1992, 81–5.

[34] JJ Cimino, SB Johnson & P Peng. Generic queries for meeting clinical information needs. Bull Med Libr Assoc 81.2 (1993), 95–206.

[35] JJ Cimino, G Elhanan & Q Zeng. Supporting infobuttons with terminological knowledge. AMIA Annual Symposium Proc. American Medical Informatics Association, 1997, 528–32.

[36] JJ Cimino, SA Socratous & PD Clayton. Internet as clinical information system: application development using the World Wide Web. Journal of the American Medical Informatics Association: JAMIA 2.5 (1995), 273–284.

[37] N Elhadad, K McKeown, D Kaufman & D Jordan. Facilitating physicians' access to information via tailored text summarization. AMIA Annu Symp Proc 2005, 226–30.

[38] E Monteiro, F Valente, C Costa & JL Oliveira. A recommender system for medical imaging diagnostic. Studies in health technology and informatics 210 (2015), 461–3.

[39] J Lin & D Demner-Fushman. Automatically evaluating answers to de nition questions. Human Language Technology and Empirical Methods in Natural Language Proc. Morristown, NJ, USA: Association for Computational Linguistics, 2005, 931–938.

[40] D Demner-Fushman & J Lin. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics Proc. 2006, 841–848.

[41] TR Goodwin & SM Harabagiu. Medical Question Answering for Clinical Decision Support. ACM International Conference on Information & Knowledge Management 2016 (2016), 297–306.

[42] I Donoso-Guzmán & D Parra. An Interactive Relevance Feedback Interface for Evidence-Based Health Care. In: Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18. New York, New York, USA: ACM Press, 2018, 103–114.

[43] JW Loonsk, R Lively, E TinHan & H Litt. Implementing the Medical Desktop: tools for the integration of independent information resources. In: Symposium on Computer Applications in Medical Care Proc 1991, 574–7.

[44] Y Yanhua Chen, L Lijun Wang, M Ming Dong & J Jing Hua. Exemplar-based Visualization of Large Document Corpus (InfoVis2009-1115). In: IEEE Transactions on Visualization and Computer Graphics 15.6 (2009), 1161–1168.

[45] D Herrmannova & P Knoth. Visual Search for Supporting Content Exploration in Large Document Collections. In: D-Lib Magazine 18.7/8 (2012).

[46] A Veerasamy & NJ Belkin. Evaluation of a Tool for Visualization of Information Retrieval Results. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '96. New York, NY, USA: ACM, 1996, 85–92.

[47] E Sherkat, S Nourashrafeddin, EE Milios & R Minghim. Interactive Document Clustering Revisited: A Visual Analytics Approach. In: 23rd International Conference on Intelligent User Interfaces - IUI '18. New York, NY, USA: ACM, 281–292.

[48] Q Zeng, JJ Cimino & KH Zou. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. In: JAMIA 9.3 (2002), 294–305.

[49] D Aleksić, P Rajković, D Vucković, D Janković & A Milenković. Data summarization method for chronic disease tracking. In: Journal of Biomedical Informatics 69 (2017), 188–202.

[50] R JL, H OM & W RA. Automating the medical record: emerging issues. In: Annu Symp Comput Appl Med Care Vol. 3. 1979, 255–263.

[51] QW O'Keefe & DW Simborg. Summary Time Oriented Record (STOR). In: Annual Symposium on Computer Application in Medical Care 2 (1980), 1175.

[52] AB Wilcox et al. Use and impact of a computer-generated patient summary worksheet for primary care. AMIA Annual Symposium Proc. Vol. 2005. American Medical Informatics Association, 2005, 824–8.

[53] MC Were et al. Creation and evaluation of EMR-based paper clinical summaries to support HIV-care in Uganda, Africa. International journal of medical informatics 79.2 (2010), 90–6.

[54] YS Lo, WS Lee, GB Chen & CT Liu. Improving the work efficiency of healthcare-associated infection surveillance using electronic medical records. Computer methods and programs in biomedicine 117.2 (2014), 351–9.

[55] BW Pickering et al. The implementation of clinician designed, human-centered electronic medical record viewer in the intensive care unit: A pilot step-wedge cluster randomized trial. International Journal of Medical Informatics 84.5 (2015), 299–307.

[56] H Liu & C Friedman. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. Studies in health technology and informatics 107.Pt 1 (2004), 639–43.

[57] SM Meystre & PJ Haug. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. Journal of Biomedical Informatics 39.6 (2006), 589–599.

[58] JA McCoy, AB McCoy, A Wright & DF Sittig. Automated Inference of Patient Problems from Medications using NDF-RT and the SNOMED-CT CORE Problem List Subset. AMIA Annual Symposium Proc. American Medical Informatics Association, 2011.

[59] DR Radev, E Hovy & K McKeown. Introduction to the Special Issue on Summarization. Computational Linguistics 28.4 (2002), 399–408.

[60] SM Meystre & PJ Haug. Randomized controlled trial of an automated problem list with improved sensitivity. International Journal of Medical Informatics 77.9 (2008), 602–612.

[61] A Wright et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 18.6 (2011), 859–867.

[62] MV Devarakonda et al. Automated problem list generation and physicians perspective from a pilot study. International Journal of Medical Informatics 105 (2017), 121–129.

[63] S Powsner. Graphical summary of patient status. The Lancet 344.8919 (1994), 386–389.

[64] SM Powsner & ER Thfte. Summarizing clinical psychiatric data. Psychiatric Services 48.11 (1997), 1458–1461.

[65] G Kopanitsa. Evaluation Study for an ISO 13606 Archetype Based Medical Data Visualization Method. Journal of medical systems 39.8 (2015), 82.

[66] S Malik et al. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. 20th International Conference on Intelligent User Interfaces. Proc 2015, 38–49.

[67] C Plaisant, J Wu, AZ Hettinger, S Powsner & B Shneiderman. Novel user interface design for medication reconciliation: an evaluation of Twinlist. JAMIA 22.2 (2015), 340–349.

[68] W Aigner & S Miksch. CareVis: Integrated visualization of computerized protocols and temporal patient data. Arti cial Intelligence in Medicine 37.3 (2006), 203–218.

[69] T David Wang et al. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. Human Factors in Computing Systems 2008, 457–466.

[70] K Wongsuphasawat et al. LifeFlow: Visualizing an Overview of Event Sequences. Human factors in computing systems. New York, New York, USA: ACM Press, 2011, 1747.

[71] T Gschwandtner, W Aigner, K Kaiser, S Miksch & A Seyfang. CareCruiser: Exploring and visualizing plans, events, and effects interactively. IEEE Paci c Visualization Symposium IEEE, 2011, 43–50.

[72] KC Spry. An infographical approach to designing the problem list. 2nd ACM SIGHIT symposium on International health informatics Proc. New York, New York, USA: ACM Press, 2012, 791.

[73] S Lee, E Kim & KA Monsen. Public health nurse perceptions of Omaha System data visualization. International Journal of Medical Informatics 84.10 (2015), 826–834.

[74] X Zhu & JJ Cimino. Clinicians evaluation of computer-assisted medication summarization of electronic medical records. Computers in Biology and Medicine 59 (2015), 221–231.

[75] Y Shahar & C Cheng. Knowledge-based visualization of time-oriented clinical data. AMIA Annual Symposium Proc. American Medical Informatics Association, 1998, 155–9.

[76] C Hallett. Multi-modal presentation of medical histories. 8th international conference on Intelligent user interfaces Proc. 2008, 80–89.

[77] W Hsu, RK Taira, S El-Saden, H Kangarloo & AAT Bui. Context-Based Electronic Health Record: Toward Patient Speci c Healthcare. IEEE Transactions on Information Technology in Biomedicine 16.2 (2012), 228–234.

[78] JS Hirsch et al. HARVEST, a longitudinal patient record summarizer. JAMIA 22.2 (2014), 263–74.

[79] V Bashyam et al. Problem-centric Organization and Visualization of Patient Imaging and Clinical Data. RadioGraphics 29.2 (2009), 331–343.

[80] AAT Bui, DR Aberle & H Kangarloo. TimeLine: Visualizing Integrated Patient Records. IEEE Transactions on Information Technology in Biomedicine 11.4 (2007), 462–473.

[81] SB Cousins & MG Kahn. The visual display of temporal information. Arti cial Intelligence in Medicine 3.6 (1991), 341–357.

[82] R Bade, S Schlechtweg & S Miksch. Connecting Time-Oriented Data and Information to a Coherent Interactive Visualization. Human Factors in Computing Systems. 2004, 105–112.

[83] L Chittaro. Information visualization and its application to medicine. Arti cial intelligence in medicine 22.2 (2001), 81–8.

[84] K Wongsuphasawat & B Shneiderman. Finding Comparable Temporal Categorical Records: A Similarity Measure with an Interactive Visualization. IEEE Symposium on Visual Analytics Science and Technology. 2009.

[85] M Glueck et al. PhenoBlocks: Phenotype Comparison Visualizations. IEEE Transactions on Visualization and Computer Graphics 22.1 (2016), 101–110.

[86] J Hunter et al. Summarising complex ICU data in natural language. AMIA Annual Symposium Proc. Vol. 2008. American Medical Informatics Association, 2008, 323–7.

[87] H Cao, MF Chiang, JJ Cimino, C Friedman & G Hripcsak. Automatic Summarization of Patient Discharge Summaries to Create Problem Lists using Medical Language Processing. MEDINFO. 2004.

[88] I Solti et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development. AMIA Annual Symposium Proc. American Medical Informatics Association, 2008, 687–91.

[89] H Cao, M Markatou, GB Melton, MF Chiang & G Hripcsak. Mining a clinical data warehouse to discover disease- nding associations using co-occurrence statistics. AMIA Annu Symp Proc. Vol. 2005. 2005, 106–110.

[90] CH Tsou, M Devarakonda & JJ Liang. Toward Generating Domain-Speci c / Personalized Problem Lists from Electronic Medical Records. AAAI 2015 Fall Symposium Proc. 2015.

[91] TT Van Vleck & N Elhadad. Corpus-Based Problem Selection for EHR Note Summarization. AMIA Annual Symposium Proc. Vol. 2010. American Medical Informatics Association, 2010, 817–21.

[92] A Goldstein, Y Shahar, E Orenbuch & MJ Cohen. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. Artificial Intelligence in Medicine 82 (2017), 20–33.

[93] R Pivovarov et al. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 58 (2015), 156–165.

[94] CW Arnold, A Oh, S Chen & W Speier. Evaluating topic model interpretability from a primary care physician perspective. Comput Methods Programs Biomed 124 (2016), 67–75.

[95] CW Arnold, SM El-Saden, AAT Bui & R Taira. Clinical Case-based Retrieval Using Latent Topic Analysis. AMIA Annu Symp Proc (2010), 26–30.

[96] R Cohen, I Aviram, M Elhadad & N Elhadad. Redundancy-Aware Topic Modeling for Patient Record Notes. PLoS ONE 9.2 (2014). Ed. by R Khanin, e87555.

[97] C Hu et al. Computational Phenotyping via Scalable Bayesian Tensor Factorization. NIPS Workshop on Machine Learning for Health. 2015.

[98] C Arnold & W Speier. A Topic Model of Clinical Reports. SIGIR Proc. 2012.

[99] E Choi et al. Multi-layer Representation Learning for Medical Concepts. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: ACM, 2016, 1495–1504.

[100] IM Baytas et al. Patient Subtyping via Time-Aware LSTM Networks. International Conference on Knowledge Discovery and Data Mining Proc. ACM Press, 2017, 65–74.

[101] ZC Lipton, DC Kale & RC Wetzel. Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks. MLHC Proc. 2017.

[102] R Miotto, L Li, BA Kidd & JT Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 6 (2016), 26094.

[103] H Suresh, P Szolovits & M Ghassemi. The Use of Autoencoders for Discovering Patient Phenotypes. NIPS Workshop on Machine Learning for Health. 2016.

[104] B Shickel, PJ Tighe, A Bihorac & P Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics 22.5 (2018), 1589–1604.

[105] P Schulam, F Wigley & S Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[106] Z Che, S Purushotham, R Khemani & Y Liu. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. arXiv preprint (2015).

[107] B Stubbs, DC Kale & A Das. Sim•TwentyFive: an interactive visualization system for data-driven decision support. AMIA Annual Symposium Proc. Vol. 2012. American Medical Informatics Association, 2012, 891–900.

[108] K Wongsuphasawat & D Gotz. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. IEEE Transactions on Visualization and Computer Graphics 18.12 (2012), 2659–2668.

[109] A Perer & D Gotz. Data-driven exploration of care plans for patients. Human Factors in Computing Systems. New York, New York, USA: ACM Press, 2013, 439.

[110] S Guo et al. EventThread: Visual Summarization and Stage Analysis of Event Sequence Data. IEEE Transactions on Visualization and Computer Graphics 24.1 (2018), 56–65.

[111] A Perer, F Wang & J Hua. Mining and exploring care pathways from electronic medical records with visual analytics. Journal of Biomedical Informatics 56 (2015), 369–378.

[112] A Perer & F Wang. Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences. 19th international conference on Intelligent User Interfaces. 2014, 153–162.

[113] N Sultanum, M Brudno, D Wigdor & F Chevalier. More Text Please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. New York, NY, USA: ACM, 2018, 422:1–422:13.

[114] J Bernard, D Sessler, A Bannach, T May & J Kohlhammer. A visual active learning system for the assessment of patient well-being in prostate cancer research. Visual Analytics in Healthcare Workshop Proc. New York, New York, USA: ACM Press, 2015, 1–8.

[115] R Joshi & P Szolovits. Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding. AMIA Annu Symp Proc. Vol. 2012. 2012, 1276–1283.

[116] S Joshi, S Gunasekar, D Sontag & J Ghosh. Identifiable Phenotyping using Constrained Non–Negative Matrix Factorization. MLHC Proc. 2016.

[117] MV Devarakonda et al. Automated problem list generation and physicians perspective from a pilot study. International Journal of Medical Informatics 105 (2017), 121–129.

[118] A Tsoukalas, T Albertson & I Tagkopoulos. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. JMIR medical informatics 3.1 (2015), e11.

[119] B Marlin, DC Kale, RG Khemani & RC Wetzel. Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models. 2nd ACM SIGHIT International Health Informatics Symposium Proc. 2012, 389–398.

[120] P Schulam & S Saria. Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses. Journal of Machine Learning Research 17.234 (2016), 1–35.

[121] R Ranganath & DM Blei. Correlated Random Measures. Journal of the American Statistical Association 113.521 (2018), 417–430.

[122] DT Bauer, S Guerlain & PJ Brown. The design and evaluation of a graphical display for laboratory data. Journal of the American Medical Informatics Association: JAMIA 17.4 (2010), 416–424.

[123] RJ Koopman et al. A diabetes dashboard and physician ef ciency and accuracy in accessing data needed for high-quality diabetes care. Annals of Family Medicine 9.5 (2011), 398–405.

[124] Y Shahar, D Goren-Bar, D Boaz & G Tahan. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artificial Intelligence in Medicine 38.2 (2006), 115–135.

[125] G Kopanitsaac, H Veselib & V Yampolskya. Development, implementation and evaluation of an information model for archetype based user responsive medical data visualization. Journal of Biomedical Informatics 55 (2015), 196–205.

[126] L Chittaro. Visualization of Patient Data at Different Temporal Granularities on Mobile Devices. Conference on Advanced visual interfaces. 2006.

[127] F Du, B Shneiderman, C Plaisant, S Malik & A Perer. Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. IEEE Transactions on Visualization and Computer Graphics 23.6 (2017), 1636–1649.

[128] R Pivovarov & N Elhadad. Automated methods for the summarization of electronic health records: Table 1. JAMIA 22.5 (2015), 938–947.

[129] JC Ho et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inform 52 (2014), 199–211.

[130] Y Kim, J Sun, H Yu & X Jiang. Federated Tensor Factorization for Computational Phenotyping. International Conference on Knowledge Discovery and Data Mining. 2017, 887–895.

[131] D Sacha, H Senaratne, BC Kwon, G Ellis & DA Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics 22.1 (2016), 240–249.

[132] D Gotz & H Stavropoulos. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. IEEE Transactions on Visualization and Computer Graphics 20.12 (2014), 1783–1792.

[133] A Rind, T Wang & W Aigner. Interactive Information Visualization to Explore and Query Electronic Health Records. Foundations and Trends® in Human–Computer Interaction 5.3 (2013), 207–298.

[134] B Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. Proc 1996 IEEE Symposium on Visual Languages. IEEE Comput. Soc. Press, 336–343.

[135] D Shahaf, C Guestrin & E Horvitz. Metro maps of information. ACM SIGWEB Newsletter Spring (2013), 1–9.

[136] W Cui et al. TextFlow: Towards Better Understanding of Evolving Topics in Text. IEEE Transactions on Visualization and Computer Graphics 17.12 (2011).

[137] HR Warner, CM Olmsted & BD Rutherford. HELP–a program for medical decision-making. Computers and biomedical research 5.1 (1972), 65–74.

[138] EH Shortliffe & DA Lindberg. Foreword. Computer-Based Medical Consultations: Mycin. 1976, pp. xvii–xxii.

[139] GO Barnett, JJ Cimino, JA Hupp & EP Hoffer. DXplain. JAMA 258.1 (1987), 67.

[140] RA Miller, HE Pople & JD Myers. Internist-I , an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. New England Journal of Medicine 307.8 (1982), 468–476.

[141] GJ Kuperman, RM Gardner & TA Pryor. Help: a dynamic hospital information system. Secaucus, NJ: Springer-Verlag, 1991.

[142] MK Goldstein et al. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modi able decision-support system for managing hypertension in primary care. AMIA Annual Symposium Proc. American Medical Informatics Association, 2000, 300–4.

[143] JH Gennari et al. The evolution of Protégé: an environment for knowledge-based systems development. International Journal of Human-Computer Studies 58.1 (2003), 89–123.

[144] AD Jeffery, LL Novak, B Kennedy, MS Dietrich & LC Mion. Participatory design of probability-based decision support tools for in-hospital nurses. JAMIA 34.11 (2017), 493–502.

[145] K Xu et al. ECGLens: Interactive Visual Exploration of Large Scale ECG Data for Arrhythmia Detection. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. New York, NY, USA: ACM, 2018, 663:1–663:12.

[146] M Jahja & DJ Lizotte. Visualizing Clinical Signi cance with Prediction and Tolerance Regions. MLHC Proc. 2017.

[147] GC Liu et al. Data visualization for truth maintenance in clinical decision support systems. International Journal of Pediatrics and Adolescent Medicine 2.2 (2015), 64–69.

[148] C Liu, F Wang, J Hu & H Xiong. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. International Conference on Knowledge Discovery and Data Mining Proc. 2015, 705–714.

[149] RS Ledley & LB Lusted. Reasoning Foundations of Medical Diagnosis. Science 130.3366 (1959), 9–21.

[150] HR Warner, AF Toronto & LG Veasy. Experience with Baye's Theorem for Computer Diagnosis of Congenital Heart Disease. Annals New York Academy of Sciences 115 (1964), 558–67.

[151] MW Pozen, RB D'Agostino, HP Selker, PA Sytkowski & WB Hood. A Predictive Instrument to Improve Coronary-Care-Unit Admission Practices in Acute Ischemic Heart Disease. New England Journal of Medicine 310.20 (1984), 1273–1278.

[152] M Cohen, D Hudson & P Deedwania. Combining ECG analysis with clinical parameters for diagnosis of heart failure. International Conference of the IEEE Engineering in Medicine and Biology Society Proc. Vol. 1. IEEE, 1997, 50–53.

[153] PJ Lisboa & AF Taktak. The use of arti cial neural networks in decision support in cancer: A systematic review. Neural Networks 19.4 (2006), 408–415.

[154] S Saria, AK Rajani, J Gould, D Koller & AA Penn. Integration of early physiological responses predicts later illness severity in preterm infants. Science translational medicine 2.48 (2010), 48ra65.

[155] CN Yu, R Greiner, HC Lin & V Baracos. Learning Patient-Speci c Cancer Survival Distributions as a Sequence of Dependent Regressors. Advances in Neural Information Processing Systems Pro. 2011.

[156] RJ Martis et al. Automated Screening of Arrhythmia Using Wavelet Based Machine Learning Techniques. Journal of Medical Systems 36.2 (2012), 677–688.

[157] DP Wall, J Kosmicki, TF DeLuca, E Harstad & VA Fusaro. Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational Psychiatry 2.4 (2012), e100.

[158] N Douali et al. Noninvasive diagnosis of nonalcoholic steatohepatitis disease based on clinical decision support system. Studies in health technology and informatics 192 (2013), 1178.

[159] AEW Johnson, AA Kramer & GD Clifford. A New Severity of Illness Scale Using a Subset of Acute Physiology and Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy. Critical Care Medicine 41.7 (2013), 1711–1718.

[160] BH Shirts, ST Bennett & BR Jackson. Using Patients Like My Patient for Clinical Decision Support: Institution-Speci c Probability of Celiac Disease Diagnosis Using Simpli ed Near-Neighbor Classi cation. Journal of General Internal Medicine 28.12 (2013), 1565–1572.

[161] SH Huang et al. Toward personalizing treatment for depression: predicting diagnosis and severity. JAMIA 21.6 (2014), 1069–1075.

[162] S Mani et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. JAMIA 21.2 (2014), 326–336.

[163] M Thomas, KD Brabanter, JA Suykens & BD Moor. Predicting breast cancer using an expression values weighted clinical classi er. BMC Bioinformatics 15.1 (2014), 411.

[164] ST Wu, YJ Juhn, S Sohn & H Liu. Patient-level temporal aggregation for text-based asthma status ascertainment. JAMIA 21.5 (2014), 876–884.

[165] M Zieba. Service-Oriented Medical System for Supporting Decisions With Missing and Imbalanced Data. IEEE Journal of Biomedical and Health Informatics 18.5 (2014), 1533–1540.

[166] TM Dugan, S Mukhopadhyay, A Carroll & S Downs. Machine Learning Techniques for Prediction of Early Childhood Obesity. Applied Clinical Informatics 6.3 (2015), 506–520.

[167] K Dyagilev & S Saria. Learning (Predictive) Risk Scores in the Presence of Censoring due to Interventions. Machine Learning Journal (2015), 1–26.

[168] A Emad, KR Varshney, DM Malioutov, TJ Watson & R Center. Learning Interpretable Clinical Prediction Rules using Threshold Group Testing. NIPS Workshop on Machine Learning for Health. 2015.

[169] P Fraccaro et al. Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. BMC Ophthalmology 15.1 (2015), 10.

[170] F Kuusisto et al. Leveraging Expert Knowledge to Improve Machine-Learned Decision Support Systems. AMIA Joint Summits on Translational Science Proc. Vol. 2015. 2015, 87–91.

[171] H Li, X Li, X Jia, M Ramanathan & A Zhang. Bone disease prediction and phenotype discovery using feature representation over electronic health records. Conference on Bioinformatics, Computational Biology and Health Informatics - BCB. '15. New York, New York, USA: ACM Press, 2015, 212–221.

[172] Z Nie, P Gong & J Ye. Predict Risk of Relapse for Patients with Multiple Stages of Treatment of Depression. International Conference on Knowledge Discovery and Data Mining Proc. 2016, 1795–1804.

[173] N Razavian, J Marcus & D Sontag. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. MLHC Proc. 2016.

[174] DA Szlosek & JM Ferretti. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. eGEMs (Generating Evidence & Methods to improve patient outcomes) 4.3 (2016), 5.

[175] J Wiens, J Guttag & E Horvitz. Patient Risk Strati cation with Time-Varying Parameters: A Multitask Learning Approach. Journal of Machine Learning Research 17 (2016), 1–23.

[176] E Choi, MT Bahadori, L Song, WF Stewart & J Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 787–795.

[177] R Henao, JT Lu, JE Lucas, J Ferranti & L Carin. Electronic Health Record Analysis via Deep Poisson Factor Models. Journal of Machine Learning Research 17 (2016), 1–32.

[178] S Horng et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS ONE 12.4 (2017). Ed. by T Groza, e0174708.

[179] Y Ling et al. Diagnostic Inferencing via Improving Clinical Concept Extraction with Deep Reinforcement Learning: A Preliminary Study. MLHC Proc. 2017.

[180] F Ma et al. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 1903–1911.

[181] K Øyvind Mikalsen et al. Computer Methods and Programs in Biomedicine Using anchors from free text in electronic health records to diagnose postoperative delirium. Computer Methods and Programs in Biomedicine 152 (2017), 105–114.

[182] I Perros et al. SPARTan: Scalable PARAFAC2 for Large & Sparse Data. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 375–384.

[183] N Reamaroon, MW Sjoding & K Najarian. Accounting for diagnostic uncertainty when training a machine learning algorithm to detect patients with the Acute Respiratory Distress Syndrome. MLHC Proc. 2017.

[184] Q Wang et al. Multi-Modality Disease Modeling via Collective Deep Matrix Factorization. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 1155–1164.

[185] K Zheng, J Gao, KY Ngiam, BC Ooi & WLJ Yip. Resolving the Bias in Electronic Medical Records. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 2171–2180.

[186] S Fong, Y Zhang, J Fiaidhi, O Mohammed & S Mohammed. Evaluation of Stream Mining Classi ers for Real-Time Clinical Decision Support System: A Case Study of Blood Glucose Prediction in Diabetes Therapy. BioMed Research International 2013 (2013), 1–16.

[187] F Hao & RH Blair. A comparative study: classi cation vs. user-based collaborative ltering for clinical prediction. BMC Medical Research Methodology 16.1 (2016), 172.

[188] Y Yang, PA Fasching & V Tresp. Modeling Progression Free Survival in Breast Cancer with Tensorized Recurrent Neural Networks and Accelerated Failure Time Models. MLHC Proc. 2017.

[189] SY Kim et al. Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Arti cial Neural Network. Korean Journal of Radiology 12.5 (2011), 588.

[190] X Wang, D Sontag & F Wang. Unsupervised Learning of Disease Progression Models. International Conference on Knowledge Discovery and Data Mining Proc. 2014.

[191] K Kourou, TP Exarchos, KP Exarchos, MV Karamouzis & DI Fotiadis. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13 (2015), 8–17.

[192] LwH Lehman et al. A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. IEEE Journal of Biomedical and Health Informatics 19.3 (2015), 1068–1076.

[193] HM Elibol et al. Cross-Corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders. Journal of Machine Learning Research 17 (2016), 1–38.

[194] Y Luo, P Szolovits, AS Dighe & JM Baron. Using Machine Learning to Predict Laboratory Test Results. American Journal of Clinical Pathology 145.6 (2016), 778–788.

[195] A McCarthy & CKI Williams. Predicting Patient State-of-Health using Sliding Window and Recurrent Classi ers. NIPS Workshop on Machine Learning for Health. 2016.

[196] SL Bergquist, GA Brooks GabrielABrooks, NL Keating, M Beth Landrum & S Rose. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. MLHC Proc. 2017.

[197] B Conroy, M Xu-Wilson & A Rahman. Patient Similarity Using Population Statistics and Multiple Kernel Learning. MLHC Proc. 2017.

[198] J Futoma et al. An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. MLHC Proc. 2017.

[199] KT Islam, CR Shelton, JI Casse, R Wetzel & LK Whittier. Marked Point Process for Severity of Illness Assessment. MLHC Proc. 2017.

[200] S Shen et al. A Bayesian model for estimating multi-state disease progression. Computers in biology and medicine 81 (2017), 111–120.

[201] J Futoma, J Morris & J Lucas. A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics 56 (2015), 229–238.

[202] A Avati et al. Improving Palliative Care with Deep Learning. IEEE International Conference on Bioinformatics and Biomedicine. 2017.

[203] P Grnarova, F Schmidt, SL Hyland & C Eickhoff. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. arXiv preprint (2016).

[204] RA Taylor et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Academic Emergency Medicine 23.3 (2016). Ed. by A Jones, 269–278.

[205] R Howard, M Rattray, M Prosperi & A Custovic. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. Current allergy and asthma reports 15.7 (2015), 38.

[206] J Liu, CE Brodley, BC Healy & T Chitnis. Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients. Artificial Intelligence in Medicine 65.2 (2015), 79–88.

[207] AA Pourzanjani, T Bo Wu, RM Jiang, MJ Cohen & LR Petzold. Understanding Coagulopathy using Multi-view Data in the Presence of Sub-Cohorts: A Hierarchical Subspace Approach. MLHC Proc. 2017.

[208] P Ordoñez, N Schwarz, A Figueroa-Jiménez, LA Garcia-Lebron & A Roche-Lima. Learning stochastic finite-state transducer to predict individual patient outcomes. Health and technology 6.3 (2016), 239–245.

[209] Y Xu, Y Xu & S Saria. A Bayesian Nonparametric Approach for Estimating Individualized Treatment-Response Curves. MLHC Proc. 2016.

[210] KM Unertl, MB Weinger, KB Johnson & NM Lorenzi. Describing and Modeling Work flow and Information Flow in Chronic Disease Care. JAMIA 16.6 (2009), 826–836.

[211] JH Chen, MK Goldstein, SM Asch, L Mackey & RB Altman. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. JAMIA 24.3 (2016), ocw136.

[212] JH Chen & RB Altman. Data-Mining Electronic Medical Records for Clinical Order Recommendations: Wisdom of the Crowd or Tyranny of the Mob? AMIA Joint Summits on Translational Science Proc. Vol. 2015. 2015, 435–9.

[213] JH Chen, T Podchiyska & RB Altman. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. JAMIA 23.2 (2016), 339–348.

[214] A Das, L Thorbergosson, A Griogorenko, D Sontag & I Huerga. Using Machine Learning to Recommend Oncology Clinical Trials. MLHC Proc (2017).

[215] JJ Gong, T Naumann, P Szolovits & JV Guttag. Predicting Clinical Outcomes Across Changing Electronic Health Record Systems. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 1497–1505.

[216] MC Hughes et al. Prediction-Constrained Topic Models for Antidepressant Recommendation. NIPS workshop on Machine Learning for Health. 2017.

[217] JG Klann, P Szolovits, SM Downs & G Schadow. Decision support from local data: Creating adaptive order menus from past clinician behavior. Journal of Biomedical Informatics 48 (2014), 84–93.

[218] S Nemati, MM Ghassemi & GD Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. EMBC. IEEE Engineering in Medicine and Biology Society (EMBC) Vol. 2016. IEEE, 2016, 2978–2981.

[219] A Raghu et al. Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. MLHC Proc. 2017.

[220] WH Weng, M Gao, Z He, S Yan & P Szolovits. Representation and Reinforcement Learning for Personalized Glycemic Control in Septic Patients. NIPS workshop on Machine Learning for Health. 2017.

[221] F Gräber et al. Therapy Decision Support Based on Recommender System Methods. Journal of Healthcare Engineering 2017 (2017), 1–11.

[222] MC Hughes, HM Elibol, T McCoy, R Perlis & F Doshi-Velez. Supervised topic models for clinical interpretability. NIPS Workshop on Machine Learning for Health. 2016.

[223] CL Jones, SM Kakade, LW Thornblade, DR Flum & AD Flaxman. Canonical Correlation Analysis for Analyzing Sequences of Medical Billing Codes. NIPS Workshop on Machine Learning for Health. 2016.

[224] FT de Dombal, DJ Leaper, JR Staniland, AP McCann & JC Horrocks. Computer-Aided Diagnosis Of Acute Abdominal Pain. Vol. 2. BMJ, 1972.

[225] HR Warner et al. ILIAD as an Expert Consultant to Teach Differential Diagnosis. Annual Symposium on Computer Application in Medical Care (1988), 371.

[226] S Saria, D Koller & A Penn. Learning individual and population level traits from clinical temporal data. NIPS Workshop on Machine Learning for Health. 2010.

[227] Y Wang et al. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. International Conference on Knowledge Discovery and Data Mining. Proc 2015.

[228] P Schulam & R Arora. Disease Trajectory Maps. arXiv preprint (2016).

[229] Y Yang et al. Predictive Clinical Decision Support System with RNN Encoding and Tensor Decoding. arXiv preprint (2016).

[230] Y Zhang, R Chen, J Tang, WF Stewart & J Sun. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. International Conference on Knowledge Discovery and Data Mining Proc. New York, New York, USA: ACM Press, 2017, 1315–1324.

[231] E Choi et al. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. MLHC Proc. 2017.

[232] EC Lorenzi, SL Brown & K Heller. Predictive Hierarchical Clustering: Learning clusters of CPT codes for improving surgical outcomes. MLHC Proc. 2017.

[233] S Parbhoo, J Bogojeska, M Zazzi, V Roth & F Doshi-Velez. Combining Kernel and Model Based Learning for HIV Therapy Selection. AMIA Joint Summits on Translational Science Proc. Vol. 2017. American Medical Informatics Association, 2017, 239–248.

[234] M Hauskrecht, S Visweswaran, GF Cooper & G Clermont. Conditional Outlier Approac Detection of Unusual Patient Care Actions h for. Twenty-Seventh AAAI Conference on Arti cial Intelligence Proc. 2013.

[235] HR Warner. Computer–assisted medical decision-making. Academic Press, 1979.

[236] RA Miller & FE Masarie. Use of the Quick Medical Reference (QMR) program as a tool for medical education. Methods of information in medicine 28.4 (1989), 340–5.

[237] R Miotto, F Wang, S Wang, X Jiang & JT Dudley. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform (2017).

[238] H Suresh et al. Clinical Intervention Prediction and Understanding with Deep Neural Networks. MLHC Proc. 2017.

[239] HU Haq, R Ahmad & SU Hussain. Intelligent EHRs: Predicting Procedure Codes From Diagnosis Codes. NIPS workshop on Machine Learning for Health 2017.

[240] H Soleimani, J Hensman & S Saria. Scalable Joint Models for Reliable Uncertainty-Aware Event Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[241] F Du, C Plaisant, N Spring & B Shneiderman. EventAction: Visual analytics for temporal event sequence recommendation. IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 2016, 61–70.

[242] AX Garg et al. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes. JAMA 293.10 (2005), 1223.

[243] TJ Bright et al. Effect of Clinical Decision-Support Systems. Annals of Internal Medicine 157.1 (2012), 29.

[244] Z Obermeyer & EJ Emanuel. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine 375.13 (2016), 1216–1219.

[245] G Hripcsak, DJ Albers & A Perotte. Parameterizing time in electronic health record studies. Journal of the American Medical Informatics Association: JAMIA 22.4 (2015), 794–804.

[246] J Zhang, H Chu, H Hong, BA Virnig & BP Carlin. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. Statistical Methods in Medical Research 26.5 (2017), 2227–2243.

[247] C Walsh & G Hripcsak. The effects of data sources, cohort selection, and outcome de nition on a predictive model of risk of thirty-day hospital readmissions. Journal of Biomedical Informatics 52 (2014), 418–426.

[248] G Hripcsak, C Knirsch, L Zhou, A Wilcox & G Melton. Bias associated with mining electronic health records. Journal of biomedical discovery and collaboration 6 (2011), 48–52.

[249] GM Weber et al. Biases introduced by ltering electronic health records for patients with "complete data". JAMIA 24.6 (2017), 1134–1141.

[250] C Paxton, A Niculescu-Mizil & S Saria. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annual Symposium Proc. Vol. 2013. American Medical Informatics Association, 2013, 1109–15.

[251] RR Shamir, T Dolber, AM Noecker, BL Walter & CC McIntyre. Machine Learning Approach to Optimizing Combined Stimulation and Medication Therapies for Parkinson's Disease. Brain Stimulation 8.6 (2015), 1025–1032.

[252] K Wongsuphasawat et al. Visualizing Data ow Graphs of Deep Learning Models in TensorFlow. IEEE Transactions on Visualization and Computer Graphics 24.1 (2018), 1–12.

[253] A Vellido, JD Martín-Guerrero & PJG Lisboa. Making machine learning models interpretable. In Proc. European Symposium on Arti cial Neural Networks, Computational Intligence and Machine Learning. 2012.

[254] JY Mao, K Vredenburg, PW Smith & T Carey. The state of user-centered design practice. Communications of the ACM 48.3 (2005), 105–109.

[255] RM Ratwani, RJ Fairbanks, AZ Hettinger & NC Benda. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. Journal of the American Medical Informatics Association 22.6 (2015), 1179–1182.

[256] A De Vito Dabbs et al. User-Centered Design and Interactive Health Technologies for Patients. Computers, informatics, nursing : CIN 27.3 (2009), 175.

[257] T McCurdie et al. mHealth consumer apps: the case for user-centered design. Biomedical Instrumentation & Technology Suppl (2012), 49–56.

[258] HealthcareUsabilityONC Meaningful Use and Usability Testing. Text. 2013.

[259] AW Kushniruk, D Kaufman, V Patel, Y Lévesque & P Lottin. Assessment of a Computerized Patient Record System: A Cognitive Approach to Evaluating Medical Technology. M.D. Computing 13.5 (1996), 406–415.

[260] AW Kushniruk & VL Patel. Cognitive and usability engineering methods for the evaluation of clinical information systems. Journal of Biomedical Informatics 37.1 (2004), 56–76.

[261] CM Johnson, TR Johnson & J Zhang. A user-centered framework for redesigning health care interfaces. Journal of Biomedical Informatics. Human-Centered Computing in Health Information Systems. Part 1: Analysis and Design 38.1 (2005), 75–87.

[262] TB Baylis, AW Kushniruk & EM Borycki. Low-Cost Rapid Usability Testing for health information systems: is it worth the effort? Studies in Health Technology and Informatics 180 (2012), 363–367.

[263] S Anders et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. International Journal of Medical Informatics 81.12 (2012), 842–851.

[264] SB Wachter et al. The employment of an iterative design process to develop a pulmonary graphical display. Journal of the American Medical Informatics Association: JAMIA 10.4 (2003), 363–372.

[265] R Verwey et al. Technology combined with a counseling protocol to stimulate physical activity of chronically ill patients in primary care. Studies in Health Technology and Informatics 201 (2014), 264–270.

[266] T Trail-Mahan, S Heisler & M Katica. Quality Improvement Project to Improve Patient Satisfaction With Pain Management: Using Human-Centered Design. Journal of Nursing Care Quality 31.2 (2016), 105–112; 113–114.

[267] M Lin, S Heisler, L Fahey, J McGinnis & TL Whiffen. Nurse Knowledge Exchange Plus: Human-Centered Implementation for Spread and Sustainability. Joint Commission Journal on Quality and Patient Safety 41.7 (2015), 303–312.

[268] DR Luna, DA Rizzato Lede, CM Otero, MR Risk & F González Bernaldo de Quirós. User-centered design improves the usability of drug-drug interaction alerts: Experimental comparison of interfaces. Journal of Biomedical Informatics 66 (2017), 204–213.

[269] JA Osheroff. Physicians' Information Needs: Analysis of Questions Posed during Clinical Teaching. Annals of Internal Medicine 114.7 (1991), 576.

[270] PC Tang, D Fafchamps & EH Shortliffe. Traditional medical records as a source of clinical data in the outpatient setting. Proceedings of the Annual Symposium on Computer Application in Medical Care (1994), 575–579.

[271] J Wentzel et al. Antibiotic information application offers nurses quick support. American Journal of Infection Control 44.6 (2016), 677–684.

[272] PN Gorman. Information needs of physicians. Journal of the American Society for Information Science 46.10 (1995), 729–736.

[273] DA Kuipers et al. iLift: A health behavior change support system for lifting and transfer techniques to prevent lower-back injuries in healthcare. International Journal of Medical Informatics 96 (2016), 11–23.

[274] D Reichert, D Kaufman, B Bloxham, H Chase & N Elhadad. Cognitive analysis of the summarization of longitudinal patient records. AMIA Annual Symposium Proc. Vol. 2010. American Medical Informatics Association, 2010, 667–71.

[275] C Bossen & LG Jensen. How physicians 'achieve overview': a case-based study in a hospital ward. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, 2014, 257–268.

[276] TT Van Vleck, DM Stein, PD Stetson & SB Johnson. Assessing data relevance for automated generation of a clinical summary. AMIA Annual Symposium Proc. American Medical Informatics Association, 2007, 761–5.

[277] S Havre, E Hetzler, P Whitney & L Nowell. ThemeRiver: visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics 8.1 (2002), 9–20.

[278] DM Blei & JD Lafferty. A Correlated Topic Model of Science. Ann Appl Stat 1.1 (2007), 17–35.

[279] T Christensen & A Grimsmo. Instant availability of patient records, but diminished availability of patient information: A multi-method study of GP's use of electronic patient records. BMC Medical Informatics and Decision Making 8.1 (2008), 12.

[280] M CJ, C FM & W A. Use of internist's free time by ambulatory care electronic medical record systems. JAMA Intern Med 174.11 (2014), 1860–1863.

[281] H RJ. Cognitive performance-altering effects of electronic medical records: An application of the human factors paradigm for patient safety. Cogn Technol Work Online 13.1 (2011), 11–29.

[282] JC Feblowitz, A Wright, H Singh, L Samal & DF Sittig. Summarization of clinical information: A conceptual model. Journal of Biomedical Informatics 44.4 (2011), 688–699.

[283] RC Li et al. Impact of problem-based charting on the utilization and accuracy of the electronic problem list. Journal of the American Medical Informatics Association (2018).

[284] J Henderson et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). AMIA Annual Symposium Proceedings 2018 (2018), 564–573.

[285] I Perros et al. SUSTain: Scalable Unsupervised Scoring for Tensors and its Application to Phenotyping. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, 2080–2089.

[286] J Pathak, AN Kho & JC Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 20.e2 (2013), e206–e211.

[287] M Conway et al. Analyzing the Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms. AMIA Annual Symposium Proceedings 2011 (2011), 274–283.

[288] Denny Joshua C. et al. Identi cation of Genomic Predictors of Atrioventricular Conduction. Circulation 122.20 (2010), 2016–2021.

[289] S Lyalina et al. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. J Am Med Inform Assoc 20.e2 (2013), e297–e305.

[290] Y Chen et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform 55 (2015), 82–93.

[291] DG Parr. Patient Phenotyping and Early Disease Detection in Chronic Obstructive Pulmonary Disease. Proceedings of the American Thoracic Society 8.4 (2011), 338–349.

[292] CA McCarty et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics 4.1 (2011), 13.

[293] Y Chen et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc 20.e2 (2013), e253–e259.

[294] F Wei et al. TIARA: A Visual Exploratory Text Analytic System. International Conference on Knowledge Discovery and Data Mining Proc 2010, 153–162.

[295] RJ Carroll, AE Eyler & JC Denny. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2011 (2011), 189–196.

[296] KP Liao et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research* 62.8 (2010), 1120–1127.

[297] S Yu et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc* 22.5 (2015), 993–1000.

[298] DM Blei, AY Ng & MI Jordan. Latent Dirichlet Allocation. *J Mach Learn Res* 3 (2003), 993–1022.

[299] C Wang & DM Blei. Variational Inference in Nonconjugate Models. *J Mach Learn Res* 14 (2013), 1005–1031.

[300] DM Blei, A Kucukelbir & JD McAuliffe. Variational Inference: A Review for Statisticians. *J Am Stat Assoc* 112.518 (2017), 859–877.

[301] Agency for Healthcare Research and Quality. *HCUP CCS. Healthcare Cost and Utilization Project (HCUP)*. 2017.

[302] DTY Wu et al. Evaluating visual analytics for health informatics applications: a systematic review from the American Medical Informatics Association Visual Analytics Working Group Task Force on Evaluation. *Journal of the American Medical Informatics Association: JAMIA* 26.4 (2019), 314–323.

[303] OHD Sciences & Informatics. *OMOP Common Data Model, url = https://www.ohdsi.org/data-standardization/the-common-data-model/, urldate = 2020-01-20*.

[304] PW Jordan, B Thomas, IL McClelland & B Weerdmeester. *Usability Evaluation In Industry*. CRC Press, 1996.