Uncovering strategies for personalized treatment selection using large language models

by

Brenda Miao

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

Atul Butte

51E283C7897F40E...

Atul Butte

Chair

DocuSigned by:

Ahmed Alaa

DA1DB75B2F0F4F5...

Ahmed Alaa

DocuSigned by:

Vivek Rudrapatna

FC4016DD9F734A9...

Vivek Rudrapatna

Committee Members

Acknowledgements

These acknowledgements cannot convey how grateful I am to everyone who has supported me throughout this journey. This thesis would not have been possible without you.

I would first like to thank my advisor, Atul Butte, whose continuous mentorship and encouragement has made me into a stronger scientist, always with real-world impact in mind. I am deeply grateful for all his insights on conducting effective and impactful research, and continue to be inspired by his unwavering optimism and dedication towards improving patient care and clinical workflows, which I hope to carry with me in all my future endeavors.

I would also like to thank all the incredible people in the Butte Lab who I have been fortunate to work with, especially Marie Mifsud and Chris Williams for all their invaluable suggestions, enthusiastic support, and for making it fun to work in person in an otherwise remote lab. Madhumita Sushil, Travis Zack, and Michelle Wang have also provided me with tremendous mentorship and collaboration on numerous occasions, as well as Zicheng Hu, who mentored me as a rotation student still working on the immunology side of the lab. And of course my fellow graduate students in the lab, Jayson and Harry, who have been a constant source of support as we progressed through graduate school together.

Thank you to the other members of my thesis committee, Vivek Rudrapatna and Ahmed Alaa, for providing guidance and perspective on these complex clinical and computational problems, and my qualifying exam committee, Tony Capra, Julian Hong, and Yulin Hswen, for taking the time to help me shape my research questions and approaches earlier in my graduate studies. I also deeply appreciate the mentorship I received from Irene Chen and Marina Sirota, whose labs I collaborated closely with and who have provided me with generous feedback and advice on my work.

I also want to thank my former research mentors Mark Anderson, Joe DeRisi, Sara Vazquez at UCSF; William Rigby and Jonathan Jones at Dartmouth College; and Li Lin and Han-sol Park at the NIH who all gave me opportunities to work in computational immunology research and led me towards my current research path. I am also grateful to have had many mentors at SandboxAQ as well, including Stefan, Sparkle, Nadia, Adam, Jordan, Constantijn, Brian, Dan, and Andrea, who gave me the opportunity to explore and shape my computational research interests in industry parallel to my academic path.

Outside of research mentors and collaborators, I am lucky to have the support of my friends Heather, Joe, Josephine, Jian, Alice, and Arvind, as we all found ourselves in various forms of graduate school these past few years. I am also grateful to have met wonderful friends at UCSF in my cohort and in the BMI program, and truly appreciate all the discussions, gatherings, and adventures we have had together.

Finally, none of this would have been possible without the support of my family, including my parents, my sister Claire, as well as Nancy, Richard, Edrea, Jerry, Michelle, and Chris, who inspire me with all that they do.

Contributions

This dissertation was supervised by Dr. Atul J. Butte.

Chapter 2 contains material originally published in The Lancet Digital Health and is available under a Creative Commons license: *Miao BY, et al. "Characterisation of digital therapeutic clinical trials: a systematic review with natural language processing." The Lancet Digital Health. 2024 Mar 1;6(3):e222-9.*

Chapter 4 contains material from a manuscript currently under review, available as an open access preprint: *Miao BY, et al. Identifying Reasons for Contraceptive Switching from Real-World Data Using Large Language Models. arXiv preprint arXiv:2402.03597. 2024 Feb 6.*

Chapter 6 contains material available through an open access preprint: *Miao BY, et al. Generation of guideline-based clinical decision trees in oncology using large language models. medRxiv. 2024:2024-03.*

Chapter 7 contains material available through an open access preprint: *Miao BY, et al. Updating the Minimum Information about CLinical Artificial Intelligence (MI-CLAIM) checklist for generative modeling research. arXiv preprint arXiv:2403.02558. 2024 Mar 5.*

Chapter 8 contains portions of text from a manuscript originally published in NPJ Digital Medicine and is available under a Creative Commons License: *Miao BY*, Mehandru N*, Almaraz ER*, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. npj Digital Medicine. 2024 Apr 3;7(1):84. *indicates co-first authorship.*

Uncovering strategies for personalized treatment selection using large language models

Brenda Miao

Abstract

Healthcare data has never been so accessible to patients and physicians, from smartphones and other remote monitoring devices to improved access for patients to their own Electronic Medical Record (EMR) history and clinical notes. Despite the ubiquity of healthcare data collection and distribution, there remains a significant gap in understanding the impacts of this data on clinical care. Insights from these digital health tools and downstream clinical decision-making processes are often only captured in medical notes, which are complex, sparse, unstructured, and difficult to model even with traditional deep learning methods. Only recently have large language models (LLMs) emerged that are capable of zero- or few-shot clinical language, without the need for large, manually annotated datasets. In this dissertation, I develop methods to adapt LLMs to healthcare tasks, particularly for identifying points of actionable insights for both digital and pharmaceutical therapeutics. These approaches demonstrate the ways in which digital health products can impact clinical care, as well as provide methods to identify reasons for medication class switching that consider the complexities of patient care beyond lab values and diagnosis codes. While careful, rigorous research is needed to ensure that these approaches are effective in facilitating patient care improvements and to reduce any potential for harm, the rapid pace of language model development provides an extraordinary opportunity to transform clinical practice. These new methods allow us to take an unprecedented look at the conversations, decisions, and medical expertise captured in billions of clinical notes and other clinical text, and to learn from this shared knowledge to accelerate medical research, improve clinical guidelines, and personalize patient care.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADI: Area Deprivation Index

AML: Acute Myeloid Leukemia

API: Application Programming Interface

BERT: Bidirectional Encoder Representations from Transformers

BoW: Bag of Words

CRC: Colorectal Cancer

DTx: Digital Therapeutic

EMR: Electronic Medical Record

FDA: US Food and Drug Administration

GPT: Generative Pre-trained Transformer

IDC: Invasive Ductal Carcinoma

IBD: Inflammatory Bowel Disease

IUD: Intrauterine Device

LDA: Latent Dirichlet Allocation

LLM: Large Language Model

MeSH: Medical Subject Headings

MI-CLAIM: Minimum Information about CLinical Artificial Intelligence

NLP: Natural Language Processing

OSCE: Objective Structured Clinical Examination

RA: Rheumatoid Arthritis

TNFα-i: Tumor Necrosis Factor Alpha Inhibitor

TF-IDF: Term-Frequency Inverse Document Frequency

UMAP: Uniform Manifold Approximation and Projection

UCSF: University of California, San Francisco

VLM: Visual Language Model

Chapter 1

Introduction

1.1 Overview

This chapter provides an overview of the dissertation, a description of the research problem, and a summary of each chapter that follows.

1.2 The Problem

Healthcare data has never been so accessible to patients and physicians. From smartphones to wearables, patients are collecting and using data in their daily lives outside the hospital, empowering them to better understand the state of their health and adjust their behaviors or seek clinical care earlier. These developments to digital devices come in parallel to improved access for patients to their Electronic Medical Record (EMR) data as well, with federal mandates for hospitals to provide digital access for patients to their medical history and clinical notes. Despite these improvements to digital health and data access, there remains a significant gap in understanding the impacts of this data on clinical care and transforming such insights into actionable clinical guidelines.

However, the nuances of patient-physician clinical decision making and clinical outcomes are often only documented in clinical notes or other real-world clinical text. Traditional natural language processing algorithms have been difficult to adapt to the clinical domain due to a lack of research datasets and specialized terminologies used. These methods relied on large, manually annotated datasets that were time-consuming and required significant clinical expertise to develop and could not be easily adapted to new tasks once trained.

The recent emergence of large language models (LLMs) has enabled more rapid clinical language modeling without the need for large, expertly-annotated training datasets. The following chapters dive into the capabilities of LLMs to understand clinical decision-making across digital and pharmaceutical therapeutic classes. The work spans several real-world datasets, starting with clinical trial registry data, electronic medical record data and clinical notes, and finally concluding with LLM capabilities on generating decision trees from best-practice clinical guidelines.

1.3 Chapters

Chapter 2 “The Digital Therapeutic Clinical Trial Landscape" describes the current landscape of FDA-regulated digital therapeutics using the ClinicalTrials.gov clinical trials registry using traditional natural language processing approaches.

Chapter 3 “Impacts of Digital Health on Clinical Care" begins to bridge the gap between digital health data and clinical care and explores the ability of the Generative Pretrained Transformer 4 (GPT-4) LLM to uncover the impacts of digital health usage from clinical notes.

Chapter 4 “Quantifying Clinical-Decision Making Using Large Language Models" assesses the use of structured medical record data as weak labels for medication information extractionand demonstrates that zero-shot GPT-4 outperforms transformer-based baseline models trained on these weak labels in extracting contraceptive prescription information. This chapter also presents results assessing the ability for LLMs to extract reasons for contraceptive switching, a novel task for LLMs.

Chapter 5 “Extracting Biologic Treatment Strategies Using Open-Source Language Models" applies the methods developed in the previous chapter to a more complex patient cohort to analyze reasons for why patients switch between tumor necrosis factor inhibitors, which are biologic therapies approved for treatment of various autoimmune diseases, and extends the previously developed medication switching information extraction pipeline to benchmark the abilities of open-source language models on this task.

Chapter 6 “Generation of Guideline-Based Clinical Decision Trees” evaluates the ability for LLMs to extract decision trees from clinical guidelines and real-world clinical notes. While previous versions of GPT, including GPT-3.5, and open-source models are not yet capable of accurate clinical decision tree extraction, we find that GPT-4 does show this emergent property, particularly with in-context guidelines provided.

Chapter 7 “Checklist for Generative Modeling for Clinical Research” provides standardized guidelines for robust study design, alignment to ethical standards, and end-to-end reproducibility in generative clinical modeling research.

Chapter 8 “Conclusions” provides concluding thoughts and a discussion of future research directions in this rapidly evolving field.

Chapter 2

The Digital Therapeutic Clinical Trial Landscape

2.1 Abstract

Digital therapeutics (DTx) are a novel class of FDA-regulated software that help patients prevent, manage, or treat disease. Here, we use natural language processing (NLP) to characterize registered DTx clinical trials and provide insights into the clinical development landscape for these novel therapeutics. We identified 449 DTx clinical trials initiated between 2010 and 2030 from ClinicalTrials.gov using 27 search terms, and available data were analyzed trial durations, locations, Medical Subject Headings (MeSH) category, enrollment, and sponsor types. Topic modeling of eligibility criteria, performed using BERTopic, showed that DTx trials frequently exclude patients based on age, comorbidities, pregnancy, language barriers, and digital determinants of health, including smartphone or data plan access. Our comprehensive overview of the DTx development landscape highlights unique challenges in designing inclusive DTx clinical and present important opportunities for clinicians and researchers to address. Finally, we provide an interactive dashboard for readers to conduct their own analyses.

2.2 Introduction

Digital therapeutics (DTx) are a novel class of FDA-regulated therapeutics that use software to help patients prevent, manage, or treat disease. Beyond providing additional therapeutic options for patients, the unique modality of DTx also enables the delivery of continuous and personalized care at scale.[1,2] Examples of approved DTx include the Propellor platform, which uses smart devices and paired consumer apps to improve medication adherence and reduces hospitalizations in patients with asthma and COPD,[3,32] and EndeavorRx, a video game that helps improve attention function in children with ADHD.[4] While DTx have the potential to help bridge gaps in access to care, there are also concerns that these software will require access to compatible devices or high digital literacy, and widen disparities in health outcomes.[1,5] These concerns prompt significant interest from healthcare and regulatory institutions to analyze the clinical development landscape and quality of clinical evidence available for DTx. [5,6]

ClinicalTrials.gov is the main site in the United States for registering clinical trials,[7] as required by Food and Drug Administration Amendments Act of 2007.[31] Several studies have previously used the ClinicalTrials.gov registry to characterize the level of clinical evidence for drug therapeutics, including analysis of clinical trial design and applicability of trial results to real-world populations.[8–10] Analogous studies of clinical trials involving digital interventions[11–13] have focused on structured data fields, and only a few have attempted to provide additional insights through manual free text analysis. However, manual analysis is time-consuming, requires specialized expertise, and difficult to keep up to date with new DTx trials. Automated tools are necessary to provide real-time insight into emerging trials from this therapeutic class.

In the last five years, developments in natural language processing (NLP) have made automated information extraction readily available for biomedical text. Software tools like SciSpacy provide open-source access to text analysis pipelines and NLP models, which are pre-trained on large biomedical datasets and can achieve high accuracies on entity extraction and other language tasks.[14,15] These pipelines can also map extracted concepts to existing biomedical vocabularies, such as Medical Subject Headings (MeSH),[16] for standardization and downstream analysis. Several NLP methods have been applied to analyze drug therapeutic clinical trials,[10,17] but have not yet been used to characterize the clinical development of DTx.

Given the increasing availability of DTx and their corresponding clinical trials we undertook a study to describe the characteristics of trials in this space. Here, we take advantage of modern NLP methods to better understand the characteristics of DTx clinical trials and the quality of evidence available for these novel therapeutics. Finally, we provide an interactive dashboard for readers to undertake their own analyses of DTx studies using both structured and unstructured data fields from ClinicalTrials.gov.

2.3 Methods

2.3.1 Search strategy and selection criteria

Digital therapeutics clinical trials were identified through the ClinicalTrials.gov application programming interface (API) using a set of 27 search terms related to DTx, including “digital therapeutic”, “digital therapy”, “smartphone”, “mobile app”, and “video game” (**Supplemental Table 2.1**). Searches were limited to the fields for BriefSummary, BriefTitle, InterventionName, InterventionDescription, Keyword, DetailedDescription, EligibilityCriteria, or OfficialTitle, and only trials registered for FDA regulated devices and not listed as having a "Basic Science” purpose were included. We use the ClinicalTrials.gov field labeled “IsFDARegulatedDevice” to identify trials that are *“studying a device product subject to section 510(k), 515, or 520(m) of the Federal Food, Drug, and Cosmetic Act*”[18] Thus, even if FDA clearance or approval has not been granted for any of these trials, this provides a high degree of confidence that these are trials for FDA-regulated products. Basic science studies were identified using the “DesignStudyPurpose” field and were removed to focus on trials of DTx with an established mechanism of action. Using the “OverallStatus” field, trials that had been terminated, withdrawn, suspended, or had an unknown status were also excluded to limit analysis to active trials. The scope of the study was also limited to studies with start dates occurring after 2010 or expected completion dates listed before 2030. Following these filtering steps, the full record from each remaining DTx trial was then extracted from the complete ClinicalTrials.gov dataset,[7] which was downloaded on August 3, 2022. We report our findings in line with PRISMA guidelines. Since this review does not assess health outcomes, no protocol is registered on PROSPERO. The full list of data fields available can be found on ClinicalTrials.gov on the Protocol Registration Data Element Definitions page.[18]

2.3.2 Analysis of clinical trial characteristics using structured data fields

The number and duration of interventional and observational trials were compared, with duration calculated in years between reported start and completion dates. Clinical trials were also analyzed based on sponsor and collaborator types, visualized using a Sankey diagram.[19] To understand the geographic distribution of clinical trials facilities for trials conducted in the United States, each entry in the LocationState field was mapped to a state code using the pgeocode software package (version 0.3.0) and the number of trials in each state was plotted as a choropleth.[19,20] The density of clinical trial facilities in each state was calculated as a ratio of trial locations to the population of each state, based on 2021 estimated US Census Bureau values. [21]

We also analyzed correlation between the number of clinical trial locations and area deprivation index (ADI), a metric of socioeconomic status in each region. ADI for the 5 states with the highest number of clinical trial locations - California, Florida, New York, Pennsylvania, and Texas - were downloaded from the University of Madison Neighborhood Atlas and mapped to each listed facility's zip code. [22,33] Both national and state level ADI were analyzed, with national ADI score given as a percentile across the entire country. At the state level, ADI is provided on a scale from 1-10. Higher scores represent greater socioeconomic disadvantage for both state and national ADIs. Only trials with available features in each data field were considered for these analyses (**Supplemental Figure 2.1**).

2.3.3 Extraction of condition and eligibility criteria using natural language processing

While ClinicalTrials.gov has an internal algorithm to map conditions listed using standardized biomedical vocabulary to MeSH terms, these terms do not correspond to the main MeSH branches and are not available for all clinical trials.[23] To create standardized mappings for each clinical trial, medical conditions from the "Condition" free text field were extracted and mapped to MeSH terms using the MeSH EntityLinker from SciSpacy,[14] with only the first match selected for each condition. Resulting terms were grouped into MeSH headings, and the most frequent heading was selected, with priority given to values under C (Diseases) and F (Psychiatry and Psychology). MeSH headings were manually reviewed for validity of the MeSH EntityLinker on this dataset.

To analyze the most common types of eligibility criteria present, we employed the BERTopic topic modeling technique,[24] which clusters text embeddings to produce interpretable, semantically cohesive clusters. BERTopic has been used in previous studies of biomedical text, and has been shown to generate more coherent topics compared to Latent Derelict Aldrich or other topic modeling methods.[25] To generate embeddings for BERTopic, text from the "Eligibility Criteria" field was first split into inclusion and exclusion criteria, with each line considered a separate document. A language model from SciSpacy pretrained on biomedical text (en_core_sci_lg) was then used to generate embeddings were generated for each eligibility criteria. A BERTopic model with default settings was used to generate topics from these embeddings, and the top five topics for each eligibility criteria were mapped back to the corresponding clinical trial to analyze the percentage of each topic occurring in each MeSH cluster. Again, a subset of the 200 inclusion criteria and exclusion criteria were manually reviewed to confirm that the eligibility criteria were mapped correctly to these topics. Only groups with at least 15 studies were analyzed.

2.3.4 Development of interactive dashboard for DTx clinical trial analysis

The dashboard for clinical trials data analysis was built using Streamlit.[26] The dashboard implements all the methods described in this paper for analysis of study types, sponsor types, conditions, and eligibility criteria. Options are available to filter the data by different study fields, and the processed dataset can be downloaded for further analysis. Our dashboard is available at https://github.com/BMiao10/ClinicalTrials.

2.3.5 Statistics

Descriptive statistics are provided for categorical variables as proportions, and averages are reported for continuous variables as medians and interquartile ranges. Spearman r values were calculated to analyze the correlation between continuous variables. Mann-U Whitney tests were used to determine differences in median enrollment between MeSH categories. Bonferroni correction was used to account for multiple testing. Statistical testing was performed using Scipy and p-values less than 0·05 were considered significant. [27]

2.4 Results

2.4.1 Identification of DTx clinical trials from ClinicalTrials.gov

Using 27 search terms related to digital therapeutics (**Supplemental Table 2.1**), we identified 8615 clinical trials involving digital-based interventions. Of these trials, 7386 were active or ongoing, and 7221 had a start date after 2010 and expected completion date before 2030. Since

DTx are regulated by the FDA as "Software as a Medical Device," we only considered studies that were listed as using FDA-regulated devices and conducted for non-basic science purposes, resulting in 449 studies of interest (**Figure 2.1**). Of these 449, there were 53 (11·8%) observational and 396 (88·2%) interventional studies (**Figure 2.2**), with 74 interventional studies listing a completion date in 2022 and 88 in 2023. Overall, 150 interventional and 18 observational studies were listed as completed, with median study durations of 1·02 years (IQR: 0·57-1·69, range: 0·06-5·17) and 0·69 years (IQR: 0·32-1·59, range: 0·05-5·42), respectively (**Figure 2.2**). When looking at dates when studies were first posted to the registry, 13 observational and 68 interventional studies were posted in 2022 (**Supplemental Figure 2.2**). Because all information on ClinicalTrials.gov is voluntarily reported by the sponsor of each clinical trial, only available data is used for each analysis and missingness is reported (**Supplemental Figure 2.1**).

2.4.2 Clinical trial locations and sponsor types

ClinicalTrials.gov requires sponsors to list facilities in which studies are being conducted, though how this is being interpreted for DTx studies is not immediately clear. As one of the primary advantages of DTx are their abilities to deliver care remotely, we were interested in understanding the geographic distribution of physical clinical trial locations listed for these trials.

Using location data provided by each study, we found that the states with the most DTx clinical trial locations were California (n=135), New York (n=58), Florida (n=55), Pennsylvania (n=52), and Texas (n=50, **Figure 2.3**), and five states - South Dakota, Wyoming, Hawaii, Delaware, and West Virginia - had no listed DTx clinical trial locations. Overall, the mean number of locations for each completed trial was 2·33 (SD 5.75). Four trials were completed without any listed facilities. The number of clinical trial locations was strongly correlated with state population (r=0·89, p<0·001, **Supplemental Figure 2.3**).

We also analyzed whether these reported clinical trial locations included socioeconomically disadvantaged neighborhoods, with socioeconomic disadvantage measured using the area deprivation index (ADI). Within the 5 states with the largest number of clinical trial locations, the number of clinical trials was inversely correlated with both the national (r=-0·52, p<0·001) and state (r=-0·66, p=0·037) ADI (**Supplemental Figure 2.3**).

To characterize the types of sponsors and collaborators funding or supporting clinical trials for the 449 trials, we looked at the lead sponsor and collaborator classes listed by each clinical trial. The most common sponsor type was "other" (n=290, 65%), which generally referred to academic medical centers. Industry was the next most common sponsor type, with 146 (33%) trials (**Figure 2.3**). The majority of studies were performed by a single sponsor with no collaborators (n=236, 53%), 131 (29%) had one collaborator, 45 (10%) had two, and 37 (8%) had three or more. For studies with a single collaborator, 26 were sponsored by other/academic institutions and had an industry collaborator, while 14 were sponsored by industry with another/academic collaborator.

2.4.3 Characteristics of digital and pharmaceutical clinical trials by medical specialty

To determine the distribution of DTx trials by medical specialty, we mapped conditions listed as free-text by each clinical trial to MeSH headings using a SciSpacy pipeline with a MeSH EntityLinker (see Methods). For trials with multiple conditions, we selected the most frequently occurring MeSH heading within that trial. The three most common MeSH headings tested in DTx clinical trials were Nervous System Diseases (n=82, 19%, **Figure 2.4**), Nutritional and Metabolic Diseases (n=45, 10%), and Pathological Conditions, Signs, and Symptoms (n=41, 9%), followed by behavior and behavior mechanisms (n=37 [8%]), cardiovascular diseases (n=34 [8%]), and mental disorders (n=31 [7%]). Examples of conditions that mapped to the heading nervous system diseases included stroke and Parkinson's disease, nutritional and metabolic diseases included both diabetes type 1 and 2, and the heading respiratory tract diseases included conditions like asthma and COPD. The MeSH category pathological conditions, signs, and symptoms contained "abnormal anatomical or physiological conditions…not classified as disease," and included conditions such as chronic pain.[16] Manual review of MeSH terms also showed that this approach mapped conditions to appropriate categories for 95% of values (**Supplemental Table 2.2**). Of the six studies in which conditions did not map to MeSH terms and were excluded from analysis, four described treatments or device characteristics (eg, device latency) rather than medical conditions and two described generic symptoms that did not map to specific headings (nasal congestion and prenatal stress, **Supplemental Table 2.3**).

With conditions classified into standardized clusters, we compared enrolment counts within each MeSH heading, focusing on non-phase 1, interventional trials in groups with fewer than ten studies. Trials targeting cardiovascular diseases had the highest number of actual and anticipated participants, with a combined median of 200 participants (IQR 100–350, range 40–450 000; 24 trials), followed by trials for nutritional and metabolic diseases with a median of 100 participants (IQR 30–197, range 6–6006; 41 trials) and behavior and behavior mechanisms again with a median of 100 participants (IQR 40–234, range 7–4500; 35 trials, **Figure 2.4**). The category with the fewest median number of participants was nervous system diseases, which had an median of 40 participants (IQR 22–100; 70 trials), although the largest trial in this category listed an anticipated enrolment of 100 000 participants. Comparing anticipated and actual enrolment information within each MeSH group, median anticipated enrolment was only significantly higher than actual enrolment for nutritional and metabolic disease DTx trials, with a median difference of 211 participants (p=0·035).

2.4.4 Extraction of eligibility criteria from digital and pharmaceutical clinical trials using natural language processing

Previous studies of drug therapeutic clinical trials have shown that eligibility criteria are often overly strict and can skew trial cohorts away from real-world patient populations.[9,10] The top five inclusion criteria topics identified by BERTopic from DTx studies were defined by terms related to clinical factors, ability to provide informed consent, age, smartphone and data access, and English fluency (**Figure 2.5**). Criteria associated with clinical factors were most frequently found in 21 (55%) of 38 pathological condition trials, 31 (47%) of 66 trials for nervous system diseases, and 11 (46%) of 24 trials for mental health disorders (**Supplemental Table 2.4**). Age criteria were most likely to be found in trials for behavioral disorders (23 [72%] of 32) and nutritional and metabolic diseases (25 [66%] of 38). Inclusion criteria detailing smartphone access were also found in several trials, occurring most frequently in DTx intended for nutritional and metabolic diseases (18 [47%] of 38) and neoplasms (8 [47%] of 17), and least frequently in trials for nervous system diseases (11 [17%] of 66) and pathological conditions (2 [5%] of 38). The topic related to smartphones and data access also contained other keywords associated with device compatibility, cellular data plans, and Wi-Fi access. Manual review of DTx studies with eligibility criteria in this topic showed patients could be excluded if they did not have a PayPal account (NCT04857515), were not willing to use a “smartphone and personal data plan” (NCT04159480) or did not show "technology literacy” (NCT04136626). This topic was found most frequently in trials for nutritional and metabolic diseases (18 [47%] of 38). Ability to provide informed consent was also most frequently found in trials for nutritional and metabolic diseases (24 [63.2%] of 38), and English fluency criteria occurred most frequently in trials for behavior and behavior mechanisms (11 [34.4%] of 32).

The top topics generated from the exclusion criteria were associated with medical history (varying between trials), pregnancy, allergies or other skin conditions, blood pressure, and, similar to inclusion criteria, ability to provide informed consent (**Figure 2.5**). There were 23 out of 24 (95·8%) DTx clinical trials targeting mental disorders, 33 out of 38 (86·8%) trials targeting nervous system diseases, and 20 out of 24 (83·3%) cardiovascular disease trials with criteria associated with medical history. Component analysis showed that some trials specifically excluded patients with smoking or suicidal behavior, cardiac disorders, or use of insulin (**Supplemental Table 2.5**). Analysis of the topic associated with pregnancy showed that nutritional and metabolic disease DTx trials were most likely to contain this exclusion criteria (n=21/38, 55·3%), while only five out of 24 trials (20·8%) for Mental Disorders and three out of 24 trials (12·5%) for Cardiovascular Diseases listed such criteria. Manual review was performed on a subset of both inclusion and exclusion eligibility criteria to ensure that topics were highly coherent and accurately described each criteria. Topics were appropriate in 94.5% (n=200) inclusion criteria and 93.5% (n=200) exclusion criteria (**Supplemental Table 2.6 – 2.7**).

2.4.5 Automated analysis of DTx data from ClinicalTrials.gov

Although ClinicalTrials.gov has filters and other data analysis tools that enable research into the structured data, there are few publicly available visual tools for the analysis of DTx clinical trials. We provide an interactive dashboard and source code for the analysis of DTx clinical trials data (**Supplemental Figure 2.4**) at https://github.com/BMiao10/ClinicalTrials.

2.5 Conclusion

Digital therapeutics are a unique modality for treating disease and have the potential to provide new treatment options for patients at an unprecedented scale. Here, we used NLP pipelines to characterize 449 DTx clinical trials identified on ClinicalTrials.gov. With more than 150 of these trials having expected completion dates by 2023, DTx are becoming rapidly available for patient care, making it essential to characterize the quality of evidence being gathered for these novel therapeutics and to better understand their benefits for real-world patient populations.

We showed that the majority of DTx trials are sponsored by academic institutions or industry with no collaborators and are primarily being developed for nervous system diseases and nutritional and metabolic diseases, which aligns with a previous review of DTx clinical trials.[13] However, the review relied on manual extraction of DTx and did not filter for FDA-regulated devices with the ClinicalTrial.gov data field. Although we were able to quantify the distribution of sponsor categories, this study did not investigate any funding sources for these sponsors or the cost of DTx trials. ClinicalTrials.gov does provide an optional field for sponsors to include information regarding grants and funding sources, but its completeness and accuracy is dependent on transparent reporting from sponsors, and future studies might be necessary to quantify funding and costs for these trials.

Our results also indicated that DTx trials were often of short duration, with interventional studies lasting an average of only 1 year, which points to a need for additional studies to understand the long-term usage and efficacy of DTx. Although these trials are short, the largest DTx trials were able to enroll more than 400 000 patients in only one or two locations, suggesting that either these trials can be effectively scaled, or that they have alternative patient recruitment strategies that ClinicalTrials.gov does not capture. However, we also showed that DTx clinical trial facilities tend to be in the most populated states. Few are done in socioeconomically disadvantaged neighborhoods, but further research is necessary to understand the true geographical and demographic distributions of users.

Analysis of DTx clinical trial eligibility criteria showed that these trials frequently exclude patients with comorbidities, who are pregnant, who are children, and who are not fluent in English. Eligibility criteria for drug therapeutics frequently cause clinical trial cohorts to deviate from real-world populations[9,10], and analogous research into DTx usage might be necessary to ensure trial results are applicable to general patient populations. We also identified criteria specific to digital determinants of health, which describe factors related to the accessibility or availability of technology that contribute to health outcomes and quality of life.[28,29] Our geographical analysis of these studies also matched this finding, which suggested that fewer facilities in disadvantaged communities in the USA are being used to recruit participants. Future initiatives to assess the role of digital determinants of health, such as SOLVE Health Tech,[30] are necessary to ensure that DTx are effective in promoting better outcomes for all patients.

The insights here and in the online interactive dashboard provide a framework for future research into DTx clinical trials, although we recognise there are limitations to our study. Although we were stringent in limiting our analysis to only FDA-regulated DTx, we might have missed DTx regulated outside the USA or inadvertently removed or selected others with our search criteria. Some DTx cleared through the 510(K) pathway, which allows medical devices to be marketed if they are "substantially equivalent" to already cleared devices, might not have registered preapproval trials,[5] but might still require post-marketing trials that could be analyzed in future studies. Additionally, we were not able to differentiate between safety and efficacy studies with the data fields provided by ClinicalTrials.gov. Our analysis is also inherently limited to sponsor-provided data, which are not always up-to-date or accurate and might be missing or unstandardised.[23] These limitations are particularly true for observational studies, for which the investigators are not required to list if they are studying an FDA-regulated product or if they accept healthy volunteers,[18] although requirements could change as regulatory pathways evolve for the use of real-world evidence in clinical trials. Finally, we focused on the use of MeSH terminology in our pipelines due to the suggested use of such terminology on ClinicalTrials.gov, but other clinical vocabularies might be more applicable to capture additional nuances in clinical trial metadata analyses. Although we took a conservative approach in mapping DTx clinical trials to broad MeSH terms, clinical trials might also involve different indications that could be better captured by allowing trials to be mapped to multiple MeSH categories.

Despite the limitations, our application of NLP strategies to ClinicalTrials.gov provides a comprehensive overview of the DTx development landscape, and the modular dashboard developed here will serve as an openly available tool for future research into clinical trial design and the real-world applicability of DTx.

2.6 Figures

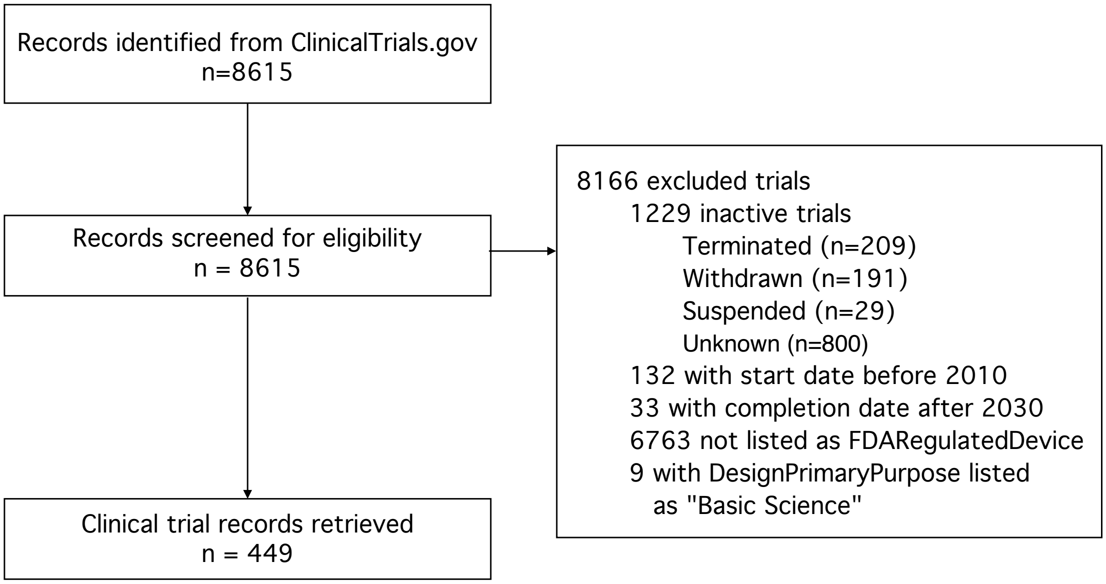

Records identified from ClinicalTrials.gov
n=8615
Records screened for eligibility
n = 8615
Clinical trial records retrieved
n = 449
8166 excluded trials
1229 inactive trials
Terminated (n=209)
Withdrawn (n=191)
Suspended (n=29)
Unknown (n=800)
132 with start date before 2010
33 with completion date after 2030
6763 not listed as FDARegulatedDevice
9 with DesignPrimaryPurpose listed
as "Basic Science"

Figure 2.1 *Identification of digital therapeutic clinical trial dataset.*

A set of 449 DTx clinical trials were identified from a query of ClinicalTrials.gov using 27 search terms and additional ClinicalTrials.gov data filters.

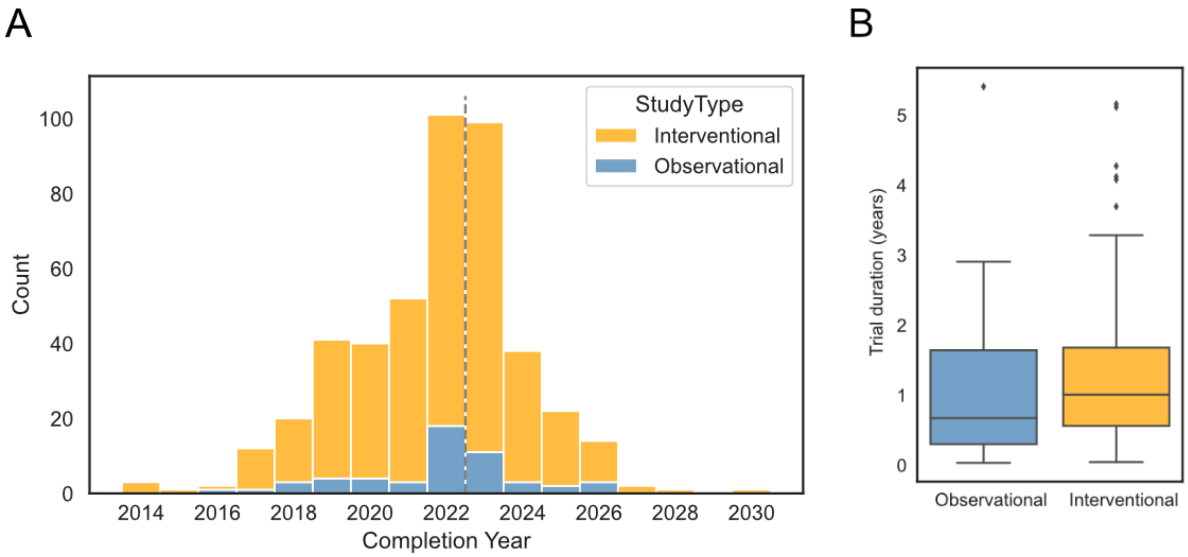


Figure 2.2 *Identification of digital therapeutic clinical trial dataset.*

A) Number of trials completed or expecting completion each year. The dashed line indicates the current year. **B)** Duration of completed interventional and observational trials.

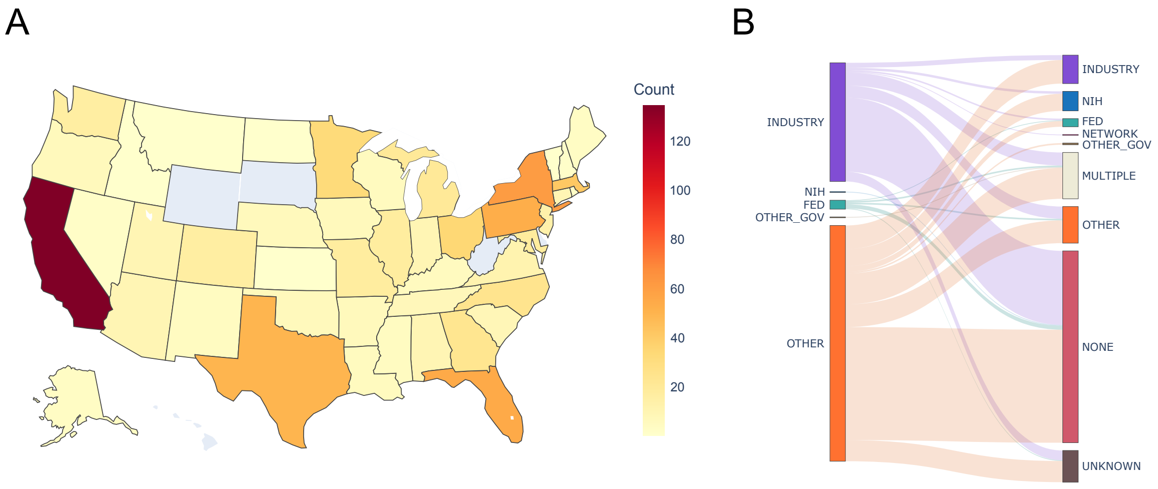


Figure 2.3 *Characteristics of US-based DTx clinical trial locations and sponsors.*

A) Number of facilities conducting DTx clinical trials by state. **B)** Distribution of sponsor (left bar) and collaborator (right bar) types.

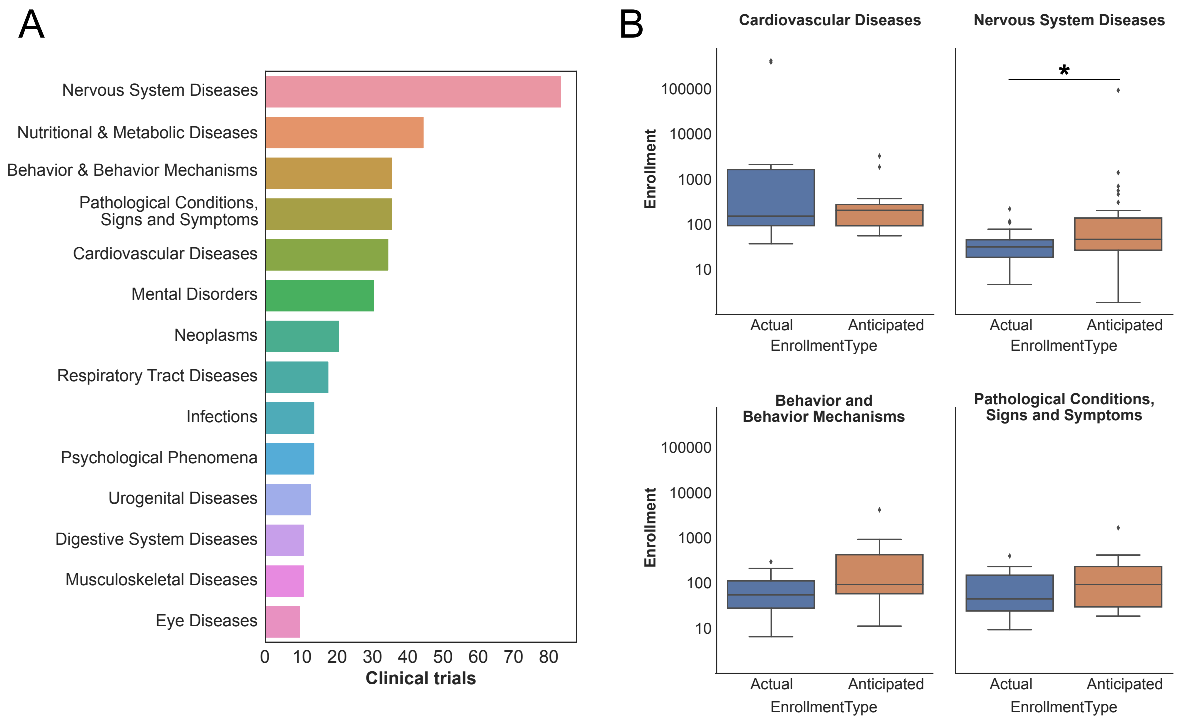


Figure 2.4 *Interventional DTx clinical trials by medical specialty.*

A) The number of clinical trials mapped to each MeSH heading using a Scispacy EntityLinker. **B)** Actual and anticipated enrollment by MeSH group.

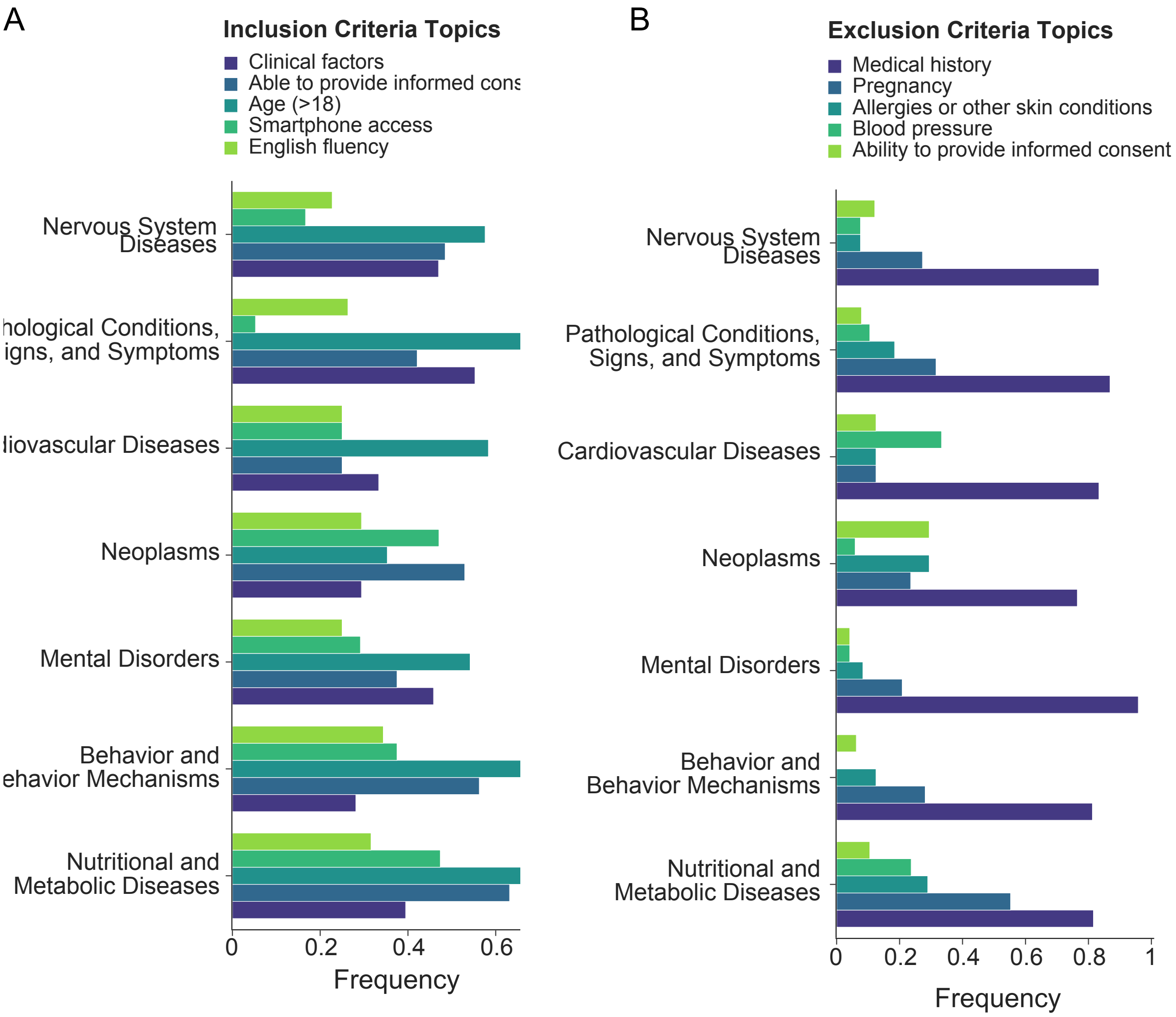


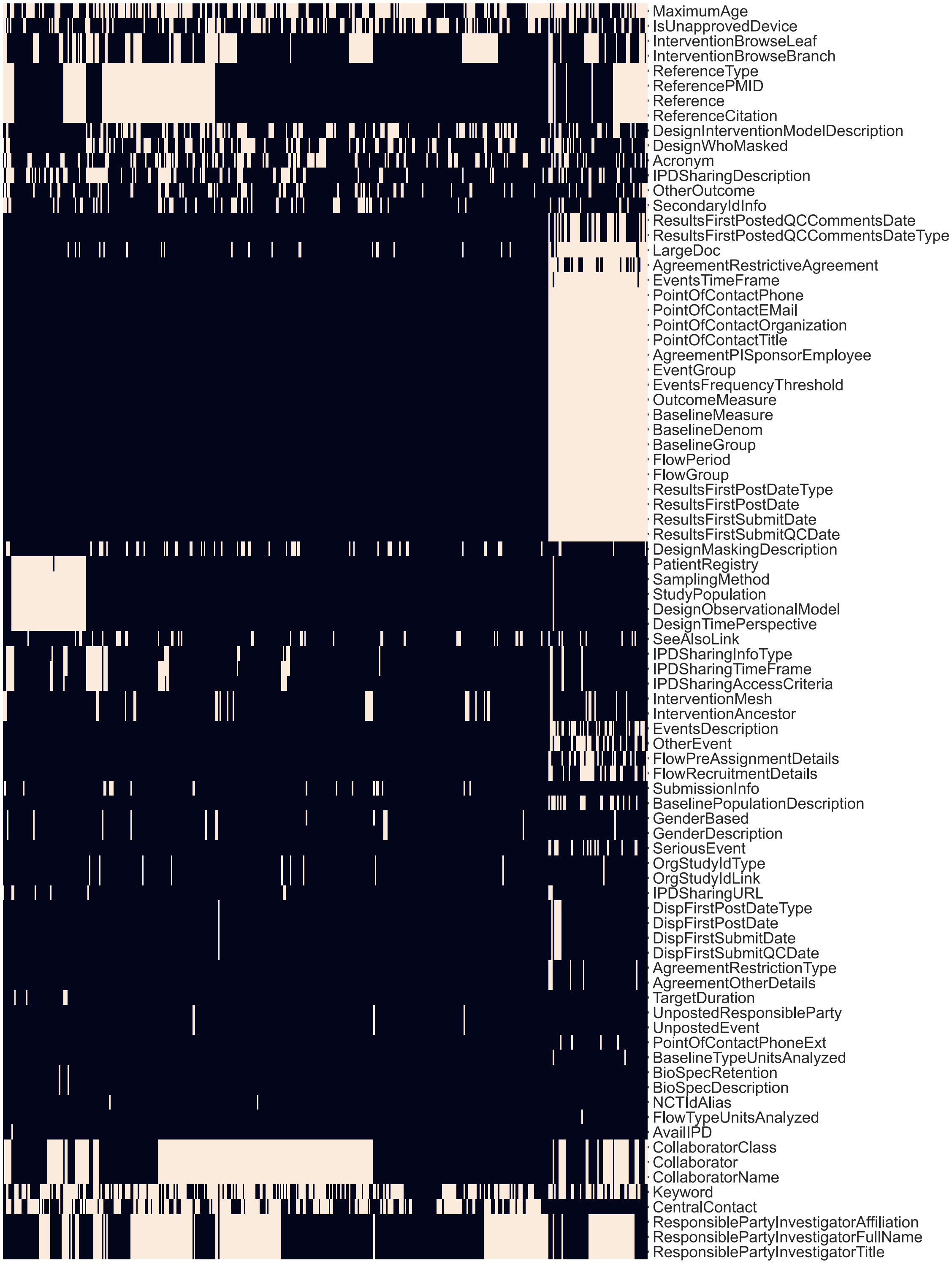
Figure 2.5 *Topic analysis of DTx clinical trial eligibility.*

BERTopic embedding clustering was used for topic modeling of **A)** inclusion and **B)** exclusion criteria of DTx trials within each MeSH heading.

2.7 Supplemental Figures and Tables

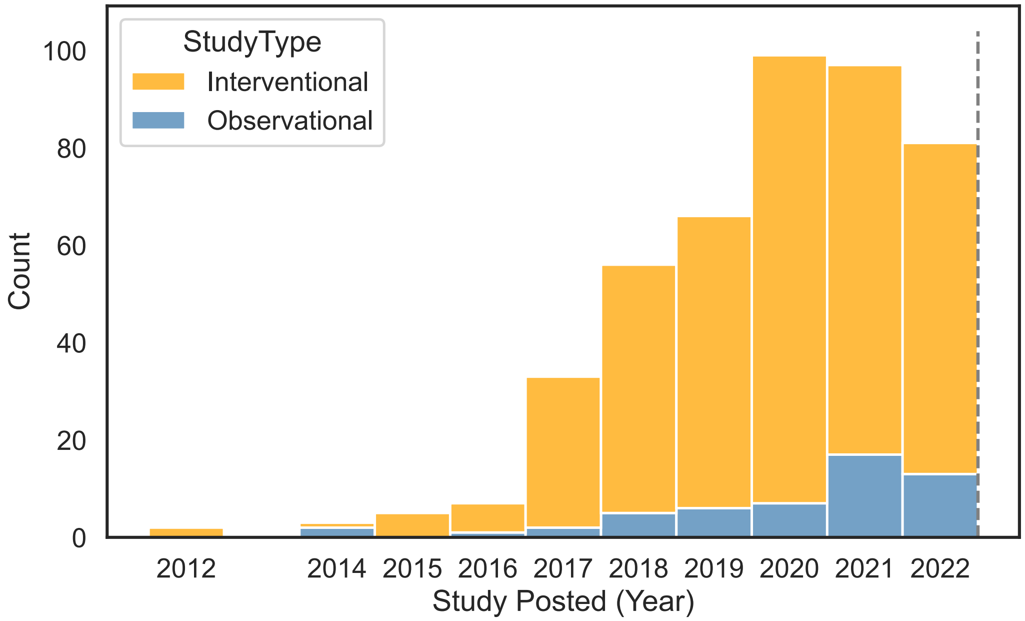
Supplemental Table 2.1 *Search results from ClinicalTrials.gov for each DTx-associated term.*

Search term	**Number of studies**
"digital therapeutic"	76
"digital therapy"	33
"digital therapies"	33
"mobile health"	1598
"smartphone"	3219
"smart phone"	3219
"digital intervention"	156
"mobile platform"	53
"mobile app"	1834
"mobile device"	462
"study app"	48
"digital treatment"	27
"android"	401
" app."	1265
" app,"	766
"digital tablet"	21
" ios"	340
"iphone"	219
"smart watch"	83
"smartwatch"	140
"virtual reality"	1747
"video game"	598
"digital health"	388
"mobile video"	12
"digital platform"	120
"software intervention"	7
"software treatment"	8

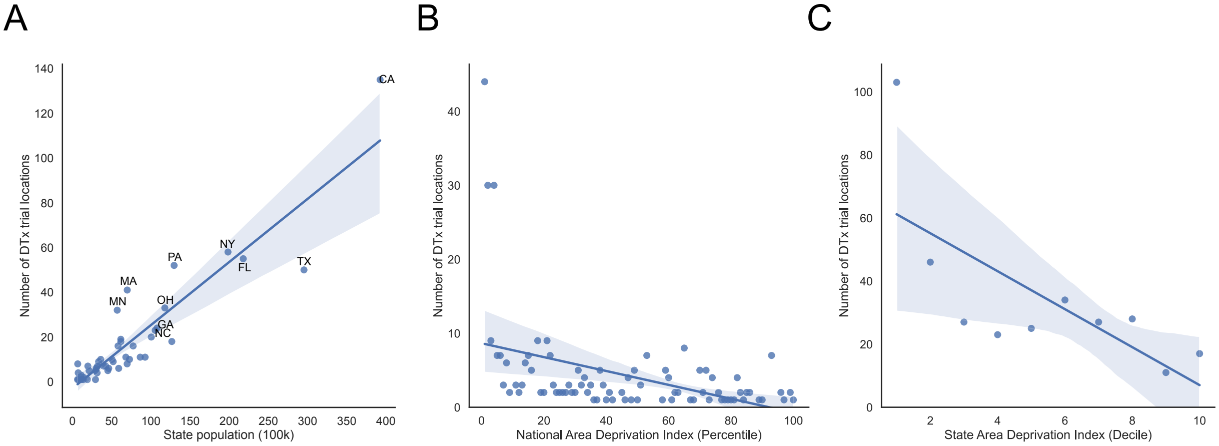


Supplemental Figure 2.1 *Missing values in DTx clinical trials for each data field.*

Only data fields where at least 1 trial contains a missing value is shown. Black bars indicate missing data.



Supplemental Figure 2.2 *Distribution of study post dates for digital therapeutic clinical trials.*



Supplemental Figure 2.3 *Correlation between DTx clinical trial locations and geographic characteristics.*

A) Correlation between number of DTx clinical trials and state population. The top 10 states with the highest number of DTx clinical trials are labeled. Within the top 5 states, we also looked at the relationship between number of DTx trial locations and area deprivation index (ADI) at the B) national level and C) state level. The national ADI is calculated relative to other zip codes in the United States, while the state ADI is relative to other zip codes in the state. Higher ADI scores at both the national and state level indicate greater socioeconomic disadvantage.

Supplemental Table 2.2 *Incorrect MeSH branches selected by SciSpacy EntityLinker.*

Clinical Trial	Condition (ClinicalTrials.gov)	MeSH branch (SciSpacy EntityLinker)	Incorrect ("N") or "Multiple" possible MeSH Branches
NCT03934658	['PostTraumatic Stress Disorder' 'Sleep Disorder' 'Stress Disorder' 'Sleep Initiation and Maintenance Disorders' 'Combat Disorders' 'Nightmares Associated With Chronic Post-Traumatic Stress Disorder' 'Nightmare' 'Nightmares, REM-Sleep Type']	Psychological Phenomena	N
NCT03828656	['PostTraumatic Stress Disorder' 'Sleep Disorder' 'Stress Disorder' 'Sleep Initiation and Maintenance Disorders' 'Combat Disorders' 'Nightmares Associated With Chronic Post-Traumatic Stress Disorder' 'Nightmare' 'Nightmares, REM-Sleep Type']	Psychological Phenomena	N
NCT03795987	['Stress Disorders, Post-Traumatic' 'Combat Disorders' 'Ptsd' 'Nightmare' 'Nightmares, REM-Sleep Type' 'Nightmare Disorder With Associated Non-Sleep Disorder']	Psychological Phenomena	N
NCT04040387	['Stress Disorders, Post-Traumatic' 'Combat Disorders' 'Ptsd' 'Nightmare' 'Nightmares, REM-Sleep Type' 'Nightmare Disorder With Associated Non-Sleep Disorder']	Psychological Phenomena	N
NCT04897074	['Attention Deficit Hyperactivity Disorder']	Psychological Phenomena	N
NCT04418076	['HIV/AIDS' 'Cocaine Use']	Organic Chemicals	N
NCT04846777	['Generalized Anxiety Disorder']	Investigative Techniques	N
NCT03748264	['Sleep Disordered Breathing']	Psychological Phenomena	N
NCT05077644	['Post-partum Depression']	Urogenital Diseases	N
NCT04364256	['Autologous Hematopoietic Stem Cell Transplant']	Biological Factors	N
NCT04684823	['Cystic Fibrosis' 'Adherence, Medication']	Digestive System Diseases	N
NCT03047720	['Nocturnal Enuresis']	Heterocyclic Compounds	N
NCT03649074	['Attention Deficit Hyperactivity Disorder']	Psychological Phenomena	N
NCT03678402	['High Risk for Falling']	Investigative Techniques	N
NCT04429009	['Thoracic Surgery' 'Respiratory Therapy']	Health Occupations	N
NCT05147987	['Insufficient Lactation']	Urogenital Diseases	N
NCT05454813	['System Validation']	Hemic and Immune Systems	N
NCT04584970	['Scoliosis Idiopathic' 'Pain, Postoperative']	Infections	N
NCT05150197	['Visual Field Defect, Peripheral']	Diagnosis	N
NCT04416555	['Postoperative Pain']	Therapeutics	N

Clinical Trial	Condition (ClinicalTrials.gov)	MeSH branch (SciSpacy EntityLinker)	Incorrect ("N") or "Multiple" possible MeSH Branches
NCT04268901	['Phlebotomy' 'Orthopedics' 'Radiology' 'Pain' 'Anxiety' 'Virtual Reality' 'Allergy' 'Gastroenterology']	Immune System Diseases	N
NCT04175444	['Visual Field Defect, Peripheral']	Diagnosis	N
NCT04025814	['Attention Deficit Hyperactivity Disorder']	Psychological Phenomena	N
NCT04857515	['Smoking Cessation' 'Smoking Behaviors' 'Smoking Reduction' 'Smoking, Cigarette' 'Smoking' 'Nicotine Dependence']	Behavior and Behavior Mechanisms	Multiple
NCT05365607	['Posttraumatic Stress Disorder' 'Cardiovascular Diseases' 'Autonomic Dysfunction' 'Vascular Stiffness' 'Nightmare' 'Endothelial Dysfunction']	Mental Disorders	Multiple
NCT03340311	['Gestational Diabetes Mellitus']	Urogenital Diseases	Multiple
NCT04808609	['Smoking Cessation' 'Smoking' 'Smoking Behaviors' 'Smoking Reduction' 'Smoking, Tobacco' 'Smoking, Cigarette' 'Hiv' 'HIV/AIDS']	Behavior and Behavior Mechanisms	Multiple
NCT04609514	['HIV/AIDS' 'Smoking Cessation' 'Tobacco Use Disorder']	Behavior and Behavior Mechanisms	Multiple
NCT04854798	['Covid19' 'Cytokine Storm' 'Inflammation']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04701489	['Covid19' 'Cytokine Storm' 'Inflammation']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04838925	['Chronic Pain' 'Opioid Use']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04217551	['Cardiac Arrest, Out-Of-Hospital' 'Hypothermia, Induced' 'Hypoxia-Ischemia, Brain']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04332718	['Stroke' 'Atrial Fibrillation']	Nervous System Diseases	Multiple
NCT03519451	['Depression' 'Tobacco Use Disorder' 'Current Every Day Smoker']	Behavior and Behavior Mechanisms	Multiple
NCT03475147	['Scotoma']	Nervous System Diseases	Multiple
NCT04465682	['Urine Detectable Acute and Chronic Diseases']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04297969	['Amblyopia Bilateral' 'Hyperopia of Both Eyes' 'Astigmatism Bilateral' 'Accommodation Disorder']	Nervous System Diseases	Multiple

Clinical Trial	Condition (ClinicalTrials.gov)	MeSH branch (SciSpacy EntityLinker)	Incorrect ("N") or "Multiple" possible MeSH Branches
NCT04607460	['Chronic Low-back Pain' 'Mastectomy' 'Lumpectomy' 'Migraine']	Nervous System Diseases	Multiple
NCT04659564	['Breast Cancer Related Lymphedema' 'Lymphedema of Upper Arm' 'Lymphedema' 'Quality of Life']	Hemic and Lymphatic Diseases	Multiple
NCT03506568	['Medication Adherence' 'Glaucoma']	Behavior and Behavior Mechanisms	Multiple
NCT04205370	['Sleep' 'Pregnancy Complications']	Psychological Phenomena	Multiple
NCT04721067	['Hiv' 'Insomnia']	Nervous System Diseases	Multiple
NCT05212129	['Functional Gastrointestinal Disorders' 'Hypermobile Ehlers-Danlos Syndrome' 'Postural Orthostatic Tachycardia Syndrome' 'Autonomic Nervous System Disease' 'Autonomic Nervous System Imbalance']	Nervous System Diseases	Multiple
NCT05099874	['Sickle Cell Disease' 'Attention Deficit' 'Cognitive Deficit in Attention']	Hemic and Lymphatic Diseases	Multiple
NCT05427734	['Suicide' 'Suicide, Attempted' 'Suicidal Ideation' 'Alcohol Use Disorder' 'Alcoholism' 'Alcohol Abuse' 'Screening and Brief Interventions']	Chemically-Induced Disorders	Multiple
NCT03905863	['Diabetic Foot Ulcer' 'Surgical Wound']	Cardiovascular Diseases	Multiple
NCT05378399	['HIV Infections' 'Substance Use' 'Adherence, Medication' 'Adherence, Treatment']	Behavior and Behavior Mechanisms	Multiple
NCT05130112	['Small Airway Disorders' 'COPD']	Pathological Conditions, Signs and Symptoms	Multiple
NCT04169282	['Tracheobronchomalacia']	Musculoskeletal Diseases	Multiple
NCT05473702	['Heart Disease Chronic' 'Pulmonary Disease, Chronic Obstructive' 'Blood Pressure' 'Heart Rhythm Disorder']	Respiratory Tract Diseases	Multiple
NCT05212363	['Compression; Vein' 'Compression; Artery' 'Sedentary Behavior' 'DVT of Legs']	Behavior and Behavior Mechanisms	Multiple

Supplemental Table 2.3 *Study conditions not mapped to MeSH by SciSpacy EntityLinker.*

Clinical Trial	Study Official Title	Condition
NCT04887922	The Effect of Preoperative and Postoperative Incentive Spirometry in Patients Undergoing Major Abdominal Surgery	Abdominal Surgery
NCT04910139	A User Study of the Soniflow System for Nasal Congestion Relief	Nasal Congestion
NCT02091882	OSMITTER 316-13-206A Substudy: A Substudy to Measure the Accuracy of Ingestible Event Marker (IEM) Detection by the Medical Information Device #1 (MIND1) System and Determine the Latency Period	Device Latency
NCT05052281	Promoting Healthy Brain Development Via Prenatal Stress Reduction: An Innovative Precision Medicine Approach	Prenatal Stress
NCT05099614	Naloxone Administration Via Auto-injection in Healthy Volunteers	Overdose Antidote
NCT05199844	Accuracy of Apple Watch to Measure Cardiovascular Indices in Patients With Cardiac Diseases: Observational Study	Apple Watch

Supplemental Table 2.4 *Keyword components comprising each inclusion criteria topic.*

Topic name	MeSH Category	Proportion	Components
Clinical factors	Pathological Conditions, Signs and Symptoms	21/38 (55·3%)	pain, months, sleep, scale, 10
	Nervous System Diseases	31/66 (47·0%)	pain, score, month, months, insomnia
	Mental Disorders	11/24 (45·8%)	treatment, medication, score, stable, month
	Nutritional and Metabolic Diseases	15/38 (39·5%)	smbg, participants, months, prior, therapy
	Cardiovascular Diseases	8/24 (33·3%)	score, states, resident, vasc, patients
	Neoplasms	5/17 (29·4%)	treatment, 20th, percentile, systemic, having
	Behavior and Behavior Mechanisms	9/32 (28·1%)	pain, usa, average, baseline, reported
Informed consent	Nutritional and Metabolic Diseases	24/38 (63·2%)	consent, informed, willing, provide, hipaa
	Behavior and Behavior Mechanisms	18/32 (56·2%)	consent, informed, provide, willing, required
	Neoplasms	9/17 (52·9%)	consent, informed, written, provide, willing
	Nervous System Diseases	32/66 (48·5%)	consent, informed, provide, willing, able
	Pathological Conditions, Signs and Symptoms	16/38 (42·1%)	consent, informed, provide, willing, signed
	Mental Disorders	9/24 (37·5%)	consent, informed, provide, willing, able
	Cardiovascular Diseases	6/24 (25·0%)	informed, consent, ascertained, able, valid
Age (>18)	Behavior and Behavior Mechanisms	23/32 (71·9%)	18, years, age, old, ages
	Nutritional and Metabolic Diseases	25/38 (65·8%)	years, age, 18, old, male
	Pathological Conditions, Signs and Symptoms	25/38 (65·8%)	18, years, age, aged, old
	Cardiovascular Diseases	14/24 (58·3%)	18, years, age, older, male
	Nervous System Diseases	38/66 (57·6%)	years, age, 18, old, aged
	Mental Disorders	13/24 (54·2%)	years, 18, age, 22, older
	Neoplasms	6/17 (35·3%)	years, 18, age, old, mole
Smartphone access	Nutritional and Metabolic Diseases	18/38 (47·4%)	smartphone, device, mobile, phone, compatible
	Neoplasms	8/17 (47·1%)	smartphone, access, recorded, audio, mobile
	Behavior and Behavior Mechanisms	12/32 (37·5%)	smartphone, android, access, iphone, ios
	Mental Disorders	7/24 (29·2%)	smartphone, iphone, game, android, controller

Topic name	MeSH Category	Proportion	Components
	Cardiovascular Diseases	6/24 (25·0%)	smartphone, plan, data, wi, fi
	Nervous System Diseases	11/66 (16·7%)	smartphone, access, device, holds, license
	Pathological Conditions, Signs and Symptoms	2/38 (5·3%)	comfortable, complete, zoom, conferencing, web
English fluency	Behavior and Behavior Mechanisms	11/32 (34·4%)	english, fluency, read, speaking, literacy
	Nutritional and Metabolic Diseases	12/38 (31·6%)	english, read, speaking, speak, write
	Neoplasms	5/17 (29·4%)	english, read, speak, speaking, write
	Pathological Conditions, Signs and Symptoms	10/38 (26·3%)	english, speaking, read, understand, spanish
	Cardiovascular Diseases	6/24 (25·0%)	english, speaking, read, write, speak
	Mental Disorders	6/24 (25·0%)	english, speaking, proficient, read, language
	Nervous System Diseases	15/66 (22·7%)	english, speaking, read, spanish, speak

Supplemental Table 2.5 *Keyword components comprising each exclusion criteria topic.*

Topic name	MeSH Category	Proportion	Components
Medical history	Mental Disorders	23/24 (95·8%)	disorder, suicidal, use, current, months
	Pathological Conditions, Signs and Symptoms	33/38 (86·8%)	disorder, history, severe, current, use
	Cardiovascular Diseases	20/24 (83·3%)	heart, weeks, patients, cardiac, disorder
	Nervous System Diseases	55/66 (83·3%)	disorder, history, severe, disease, months
	Nutritional and Metabolic Diseases	31/38 (81·6%)	insulin, disease, disorder, months, investigator
	Behavior and Behavior Mechanisms	26/32 (81·2%)	disorder, current, smoking, suicidal, treatment
	Neoplasms	13/17 (76·5%)	disorder, history, care, therapy, psychiatric
Pregnancy	Nutritional and Metabolic Diseases	21/38 (55·3%)	pregnant, pregnancy, women, female, feeding
	Pathological Conditions, Signs and Symptoms	12/38 (31·6%)	pregnant, pregnancy, teeth, women, face
	Behavior and Behavior Mechanisms	9/32 (28·1%)	pregnant, planning, face, pregnancy, comfortable
	Nervous System Diseases	18/66 (27·3%)	pregnant, pregnancy, women, breastfeeding, potential
	Neoplasms	4/17 (23·5%)	pregnant, adults, populations, vulnerable, prisoners
	Mental Disorders	5/24 (20·8%)	pregnant, pregnancy, teenagers, cycles, menstrual
	Cardiovascular Diseases	3/24 (12·5%)	postpartum, wfbmc, center, location, birth
Allergies or other skin conditions	Neoplasms	5/17 (29·4%)	cancer, documented, skin, patients, hematologic
	Nutritional and Metabolic Diseases	11/38 (28·9%)	skin, cell, allergy, basal, neoplasms
	Pathological Conditions, Signs and Symptoms	7/38 (18·4%)	skin, cancer, carcinoma, sores, hardware
	Behavior and Behavior Mechanisms	4/32 (12·5%)	skin, fragile, dermatologic, intact, oozing
	Cardiovascular Diseases	3/24 (12·5%)	wound, patches, surface, adhesive, skin
	Mental Disorders	2/24 (8·3%)	cancers, skin, cvd, preexisting, angiomas
	Nervous System Diseases	5/66 (7·6%)	allergic, skin, known, tapes, reaction
Cardiovascular metrics	Cardiovascular Diseases	8/24 (33·3%)	mmhg, diastolic, cm, baseline, dl
	Nutritional and Metabolic Diseases	9/38 (23·7%)	mmhg, pressure, 60, ml, min

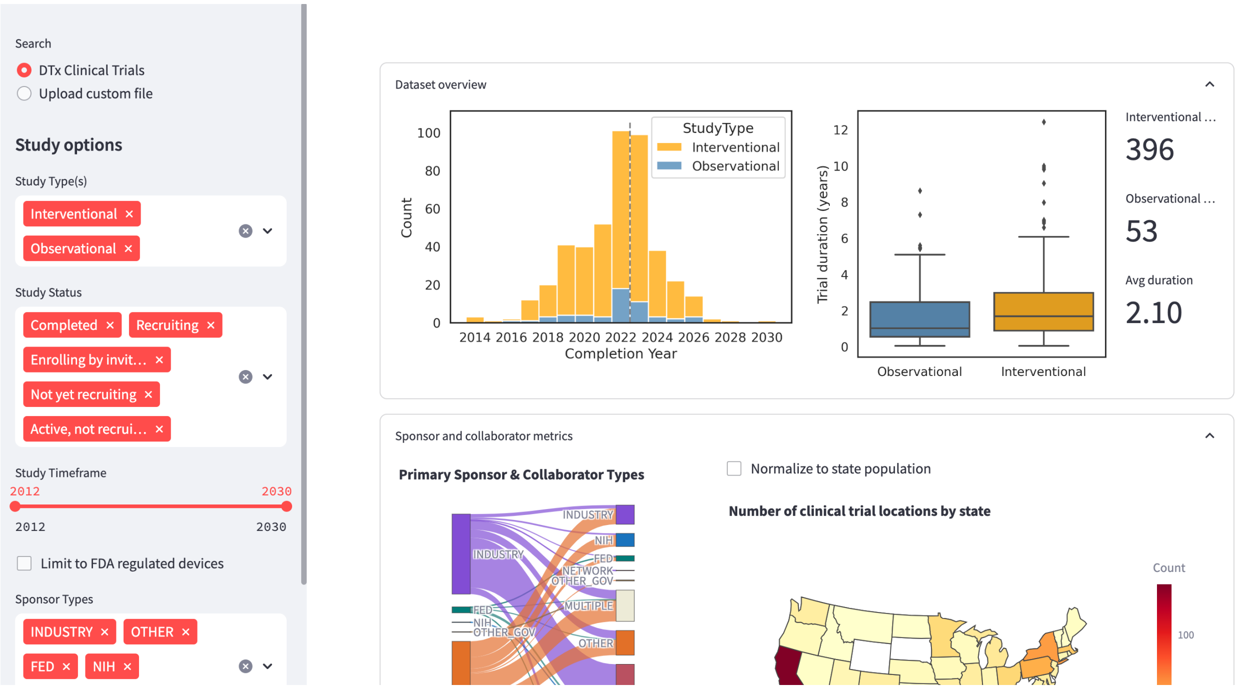
Topic name	MeSH Category	Proportion	Components
	Pathological Conditions, Signs and Symptoms	4/38 (10·5%)	inches, 79, clots, obesity, circumference
	Nervous System Diseases	5/66 (7·6%)	mmhg, 60, diastolic, 30, cm
	Neoplasms	1/17 (5·9%)	mass, 35, bmi, index, body
	Mental Disorders	1/24 (4·2%)	tsh, thyroid, pcp, mu, values
Ability to provide informed consent	Neoplasms	5/17 (29·4%)	consent, informed, sign, willing, inability
	Cardiovascular Diseases	3/24 (12·5%)	consent, provide, informed, inability, unwilling
	Nervous System Diseases	8/66 (12·1%)	consent, informed, provide, written, unable
	Nutritional and Metabolic Diseases	4/38 (10·5%)	consent, informed, inability, provide, unwillingness
	Pathological Conditions, Signs and Symptoms	3/38 (7·9%)	consent, informed, provide, inability, unable
	Behavior and Behavior Mechanisms	2/32 (6·2%)	informed, consent, inability, provide, unwillingness
	Mental Disorders	1/24 (4·2%)	consent, informed, adults, unable, written

Supplemental Table 2.6 *Incorrect values from manual assessment of topic modeling in a subset of 200 inclusion eligibility criteria.*

NCTId	Inclusion Criteria (Preprocessed)	BERTopic cluster	Expected cluster
NCT03142932	agree anticipate living baltimore 2 months	Clinical factors	Other
NCT04380415	u.s. resident	Clinical factors	Other
NCT03214224	1	Clinical factors	NA
NCT03418129	served military branches army navy marines air force coast guard october 2001.	Clinical factors	Other
NCT03335800	current resident united states time eligibility screening defined self reported state residence 50 states united states district columbia.	Clinical factors	Other
NCT04524598	residing usa duration 5 week	Clinical factors	Other
NCT04268914	4. children normal range development recruited study. assessed report parents. rationale excluding patients developmental delay cognitive impairments children react stressors surgery differently children developmental delay. unclear children use preparation programs interventions included likely responses baseline outcome measures differ children normal developmental parameters.	Clinical factors	Ability to provide informed consent
NCT04253691	unable complete forms implement treatment cognitive impairment mmse<26	Clinical factors	Ability to provide informed consent
NCT04607460	able speak understand english 6 access computer tablet home email address.	Ability to provide informed consent	English fluency
NCT03335800	valid phone number associated iphone ascertained self report.	Ability to provide informed consent	Smartphone access
NCT03338036	holds valid driver s license	Smartphone access	Other

Supplemental Table 2.*7 Incorrect values from manual assessment of topic modeling in a subset of 200 exclusion eligibility criteria.*

NCTId	Inclusion Criteria (Preprocessed)	BERTopic cluster	Expected cluster
NCT05263037	members household	Medical history	Other
NCT04394754	incarceration	Medical history	Other
NCT03528174	hematocrit 36 men 32 women.	Pregnancy	Cardiovascular metrics
NCT04479735	refuses lidocaine 2.5 prilocaine 2.5 cream use excluded study.	Medical history	Ability to provide informed consent
NCT05293275	injury eyes face neck impedes comfortable use virtual reality	Medical history	Medical history
NCT04797611	agree use approved contraception method entirety trial	Medical history	Ability to provide informed consent
NCT05263037	injury eyes face neck prevents comfortable use vr.	Pregnancy	Ability to provide informed consent
NCT04906603	loss consciousness greater 30 minutes	Cardiovascular metrics	Medical history
NCT04230486	volunteers unable complete tasks understand instructions	Medical history	Ability to provide informed consent
NCT05112432	legal commitment treatment medical guardianship provision guardianship order court order allow guardian consent research	Medical history	Ability to provide informed consent
NCT03315286	employee direct relative employee investigational site sponsor	Medical history	Ability to provide informed consent
NCT04152447	injuries requiring staged surgical fixation i.e. ex fix orif	Allergies or other skin conditions	Medical history
NCT03996954	patients unable unwilling use device	Medical history	Ability to provide informed consent



Search

● DTx Clinical Trials

○ Upload custom file

Study options

Study Type(s)

Interventional × Observational ×

Study Status

Completed × Recruiting × Enrolling by invit… × Not yet recruiting × Active, not recrui… ×

Study Timeframe

2012 2030

2012 2030

□ Limit to FDA regulated devices

Sponsor Types

INDUSTRY × OTHER × FED × NIH ×

Dataset overview

Interventional …

396

Observational …

53

Avg duration

2.10

Sponsor and collaborator metrics

Primary Sponsor & Collaborator Types

□ Normalize to state population

Number of clinical trial locations by state

***Supplemental Figure 2.4** Clinical trials analysis dashboard.*

Screenshot of an interactive dashboard for analysis of ClinicalTrials.gov metadata for DTx clinical trials, provided for readers.

References

1. Miao, B. Y., Arneson D., Wang M., & Butte, A. J. Open challenges in developing digital therapeutics in the United States. *PLOS Digit Health*. 1: e0000008 (2022).

2. Patel, N. A., Butte A. J. Characteristics and challenges of the clinical pipeline of digital therapeutics. *NPJ Digit Med*. 3: 159 (2020).

3. Moore, A. *et al*. A randomised controlled trial of the effect of a connected inhaler system on medication adherence in uncontrolled asthmatic patients. *Eur Respir J*. 57: 2003103 (2021).

4. Kollins, S. H., Childress, A., Heusser, A. C., & Lutz J. Effectiveness of a digital therapeutic as adjunct to treatment with medication in pediatric ADHD. *NPJ Digit Med*. 4: 58 (2021).

5. Torous, J., Stern, A. D., & Bourgeois, F. T. Regulatory considerations to keep pace with innovation in digital health products. *NPJ Digit Med*. 5: 121 (2022).

6. Crisafulli, S., Santoro, E., Recchia, G., & Trifirò, G. Digital therapeutics in perspective: from regulatory challenges to post-marketing surveillance. *Front Drug Saf Regul*. 2: 900946 (2022).

7. ClinicalTrials.gov. National Library of Medicine. https://clinicaltrials.gov/. Last accessed August 26, 2022.

8. Califf, R. M., Zarin, D. A., Kramer, J. M., Sherman, R. E., Aberle, L. H., & Tasneem, A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA*. 307: 1838–47 (2012).

9. Rudrapatna, V. A., Glicksberg, B. S., & Butte, A. J. A comparison of the randomized clinical trial efficacy and real-world effectiveness of tofacitinib for the treatment of inflammatory bowel disease: a cohort study. *medRxiv*. https://doi.org/10.1101/19007195 (2019). (preprint).

10. Liu, R., Rizzo, S., Whipple, S., et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 592: 629–33 (2021).

11. Marra, C., Chen, J. L., Coravos, A., & Stern, A. D. Quantifying the use of connected digital products in clinical research. *NPJ Digit Med*. 3: 50 (2020).

12. Bartlett, V. L., Dhruva, S. S., Shah, N. D., & Ross, J. S. Clinical studies sponsored by digital health companies participating in the FDA's Precertification Pilot Program: a cross-sectional analysis. *Clin Trials.* 19: 119–22 (2022).

13. Santoro, E., Boscherini, L., & Caiani, E. G. Digital therapeutics: a systematic review of clinical trials characteristics. *Eur Heart J*. 42 (suppl 1): 724–3115 (2021).

14. Neumann, M., King, D., Beltagy, I., & Ammar, W. ScispaCy: fast and robust models for biomedical natural language processing. Association for Computational Linguistics. *Proceedings of the 18th BioNLP Workshop and Shared Task*: 319–327 (2019).

15. Beltagy, I., Lo, K., & Cohan, A. SciBERT: a pretrained language model for scientific text. Association for Computational Linguistics. *EMNLP-IJCNLP*: 3615–3620 (2019).

16. US National Library of Medicine. Medical Subjects Headings: Introduction to MeSH – 2010. http://www.nlm.nih.gov/mesh/2010/introduction/introduction.html. Last accessed March 28, 2022.

17. Fang, Y., Idnay, B., Sun, Y., et al. Combining human and machine intelligence for clinical trial eligibility querying. *J Am Med Inform Assoc*. 29: 1161–71 (2022).

18. US National Library of Medicine. ClinicalTrials.gov protocol registration data element definitions for interventional and observational studies. 2020. https://prsinfo.clinicaltrials.gov/definitions.html. Last accessed August 29, 2022.

19. Plotly Technologies. Collaborative data science. Montréal, QC (2015). https://plot.ly.

20. Yurchak, R. pgeocode 0·3.0. 2020 Oct 23. https://github.com/symerio/pgeocode.

21. US Census Bureau. State population totals and components of change: 2020–2021. 2021. https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html. Last accessed August 15, 2022.

22. Kind, A. J. H., & Buckingham, W. R. Making neighborhood disadvantage metrics accessible: the Neighborhood Atlas. *New Engl J Med*. 378: 2456–58 (2018).

23. Miron, L., Gonçalves, R. S., & Musen, M. A. Obstacles to the reuse of study metadata in ClinicalTrials.gov. *Sci Data*. 7: 443 (2020).

24. Grootendorst, M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. https://arxiv.org/abs/2203.05794 (2022). (preprint).

25. Meaney, C., Escobar, M., Stukel, T. A., & Austin, P. C., Jaakkimainen, L. Comparison of methods for estimating temporal topic models from primary care clinical text data: retrospective closed cohort study. *JMIR Med Inform*. 10: e40102 (2022).

26. Streamlit. https://streamlit.io/.

27. Virtanen, P., Gommers, R., Oliphant, T. E., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17: 261–72 (2020).

28. Richardson, S., Lawrence, K., Schoenthaler, A. M., & Mann, D. A framework for digital health equity. *NPJ Digit Med*. 5: 119 (2022).

29. Sieck, C. J., Sheon, A., Ancker, J. S., Castek, J., Callahan, B., & Siefer, A. Digital inclusion as a social determinant of health. *NPJ Digit Med*. 4: 52 (2021).

30. Lyles, C. R., Sarkar, U., Patel, U., et al. Real-world insights from launching remote peer-to-peer mentoring in a safety net healthcare delivery setting. *JAMIA*. 28: 365–70 (2021).

31. Food and Drug Administration. Food and Drug Administration Amendments Act (FDAAA) of 2007. 2007. http://www.fda.gov.

32. Merchant, R., Szefler, S. J., Bender, B. G., et al. Impact of a digital health intervention on asthma resource utilization. *World Allergy Organ J.* 11: 1-4 (2018).

33. University of Wisconsin School of Medicine Public Health. About the Neighborhood Atlas: 2015 Area Deprivation Index v2·0. https://www.neighborhoodatlas.medicine.wisc.edu/ (accessed Aug 26, 2022).

Chapter 3

Impacts of Digital Health on Clinical Care

3.1 Abstract

Digital health is rapidly growing in importance in modern healthcare, but its uptake and utility across patients and clinicians remains unclear. We analyzed over 100 million clinical notes from the Information Commons dataset at the University of California, San Francisco and found more than half a million notes containing digital health terms curated from PubMed articles and ClinicalTrials.gov. We characterized patient demographics, department, and frequency over time, and trends identified using clinical language modeling approaches. We identified 209,377 notes from 66,121 unique patients containing any of 91 digital curated digital health terms. These notes were primarily written for younger, White, and English speaking patients, and most frequently documented in pediatric endocrinology, cardiology, and neurology departments. Digital health documentation grew at an average annual rate of 26.9% from 2012-2022. Topic modeling analysis identified 29 clusters in digital health documentation, with glucose monitoring and Apple Watch usage in cardiology as the most frequent topics. Information extracted using GPT-4 applied to 6,940 cardiology notes mentioning “Apple Watch” showed that 53.2% of the smartwatch usage was patient-initiated, and was primarily used for ongoing disease monitoring. In 36.8% of these notes, new patient instructions were provided, 21.2% led to procedure orders, and 15.6% led to medication changes. Our findings identified trends in digital health usage, and provide a framework for characterizing downstream impact of digital health using natural language processing approaches. Results also highlight the need for more evidence-based resources for digital health usage and access.

3.2 Introduction

Digital health refers to the use of digital technologies to improve health, healthcare, and clinical outcomes, and encompasses a wide range of hardware and software tools[1]. These tools are increasingly being used both in clinical settings and by patients outside the healthcare system to track and monitor their own health. Real-world studies of digital health through retrospective medical record data analysis or prospective clinical trials have shown that digital health can help scale monitoring and disease management efforts outside the hospital[2–4,5,6,7], streamline clinical workflows[8,9], and provide clinical decision support[10–12], and digital tools like telemedicine can often be an effective alternative to in-person visits[13–15]. Despite the potential of digital health to aid in clinical care[16], there has also been research highlighting the challenges in the implementation of digital health or disparities in access, particularly for socioeconomically disadvantaged patients and in elderly populations[17–19]. As these digital health tools are adopted into routine healthcare, identifying best-practice guidelines for their use requires a better understanding of their real-world utilization and impact.

One significant challenge to the study of digital health in real-world data is a lack of standardized ontologies and frameworks to identify digital health utilization[20–22], particularly within unstructured clinical notes. Clinical notes are a rich source of unstructured data detailing a patient’s health status and clinical treatment plans[23] and are particularly suitable for documenting aspects of patient health that occur outside the hospital, including digital health access and usage[24]. In this study, we take a data-driven approach to quantify how digital health is used and documented in real-world clinical text. This comprehensive assessment of digital health usage aims to support the development of evidence-based ontologies for shared clinical decision making around digital health usage.

3.3 Results

3.3.1 Digital health documentation dataset characterization.

We identified 91 digital health terms from a search of PubMed articles, clinical trial descriptions listed on ClinicalTrials.gov, and previously curated publications (**Supplemental Table 3.1**). Out of 136,026,361 total inpatient, outpatient, ambulatory, and ancillary notes from Information Commons (2012-2022)[23,25], we found 618,943 notes (0.46%, data retrieved July 17, 2023) containing mentions of any of these digital health terms. Following deduplication and removal of telehealth or online patient portal setup instructions, the final dataset contained 209,377 notes from 66,121 unique patients (**Figure 3.1**).

Among patients with digital health notes, 36,315 (54.9%) listed their gender as female, 29,744 (44.9%) male, and 24 (0.0%) nonbinary (**Table 3.1**). The mean age of patients at the time of digital health documentation was 37.4 years (n=66,121, SD: 24.9 years), compared to a mean of 44.1 years for patients without digital health documentation (n=2,169,595, SD: 25.8 years, p<0.001). Patients with digital health notes were also significantly more likely to list "English" as their preferred language (91.4%) compared to patients without (83.0%, p<0.001).

Distributions of self-reported race and ethnicity also differed significantly between patients with and without digital health notes (p<0.001). The largest proportion of patients with digital health notes self-reported as "White" (47.4%), which was higher than the proportion of "White" patients without digital health notes (37.4%). Similar trends were seen in patients self-reporting as "Latinx" (16.1% vs 15.1%), "Asian" (14.5% vs 10.3%), or "Black or African American" (7.8% vs 5.4%). Only 4.1% of patients with digital health notes listed their race as "Unknown/Declined" (4.2%, n=2774) compared to 21.8% (n=473,673) of patients without digital health notes.

3.3.2 Characterization of digital health documentation.

The majority of digital health notes written by physicians, residents, and nurse practitioners (**Supplemental Table 3.2**), and occurred most frequently in pediatric endocrinology (n=21746, 11.5%), cardiology (n=18966, 10.1%), neurology (n=13592, 7.2%) departments (**Figure 3.2**). The most frequently mentioned terms were “smartphone” (n=34,630, 16.5%), “FreeStyle Libre” (n=32,909, 15.7%), “iPhone” (n=31,168, 14.9%), “Contour Next” (n=23,106, 11.0%), and “Apple Watch” (n=14,319, 6.8%, **Supplemental Table 3.2**).

Distributions of the top 10 terms differed significantly between department specialties (p<0.001, **Figure 3.2**). Within the pediatric endocrinology notes, “Contour Next” (44.0%, n=7986) and “Freestyle Libre” (42.8%, n=7761) were most frequently mentioned. “Contour Next” and “Freestyle Libre” are both terms that capture connected glucose monitoring devices and associated phone applications[26,27]. In cardiology, “Remote monitoring” (35.6%, n=6514) and “Apple Watch” (39.7%, n=7270) were the most frequently used terms. “Smartphone” was the most common term occurring in both neurology (50.0%, n=6741) and primary care (42.5%, n=4044), while “iPhone” was the most frequently documented term in audiology (48.0%, n=5493).

3.3.3 Documentation of digital health in clinical notes is rapidly increasing.

The compound annual growth rate (CAGR) of digital health term documentation from 2012 to 2022 was 26.9%, compared to the growth of all clinical notes (7.31%, **Figure 3.3**). Of the top 10 most prevalent terms, the terms with the highest CAGRs were “Freestyle Libre” (147.6%, n=37317), “Apple Watch” (127.7%, n=15008), and “App-based” (88.4%, n=4481). Documentation of the terms “Smartphone” (n=37456), “Android” (n=10386), and “iPhone” (n=34077) have also been increasing, with CAGRs of 31.0%, 30.8%, and 13.0%, respectively. Out of the most prevalent terms, only the term “Internet-based” showed a decline in usage with a CAGR of -29.3% (n=4205). CAGR values for all other terms ranged from -10.4% to 82.2% (**Supplemental Table 3.2**).

The change of digital health mentions from 2012-2022 across different departments was also examined. Within departments with the greatest number of notes, the rate of growth in digital health documentation was greatest in Primary Care, with a CAGR of 50.3% (n=11049). Other departments showed similar rates of growth, with CAGRs ranging from 22.0% in General Pediatrics (n=3873) to 33.2% in Inpatient Nursing (n=6607, **Figure 3.3**). Of all departments, Psychiatry had the highest average growth in digital health documentation, with a CAGR of 89.6% (n=3110), and General Surgery had the lowest, with a negative CAGR of -15.1% (n=3538, **Supplemental Table 3.3**).

3.3.4 Context of digital health usage in clinical notes

Topic modeling using Latent Dirichlet Allocation (LDA) identified 29 clusters of digital health documentation (**Figure 3.4**, **Supplemental Table 3.4**). The most common topics within the digital health dataset were related to “Contour Next” glucose monitoring (topic 2, n=19050, **Supplemental Table 3.3**), “Apple Watch” usage in cardiology (topic 10, n=17569), medications (topic 1, n=17052), and glucose monitoring usage (topic 5, n=15105, **Figure 3.4**). Manually curated examples of notes from these categories are shown in **Figure 3.5**. Other topics contained terms describing access to devices (topic 21, n=13370), iPhones and hearing aid use (topic 17, n=12038), mobile apps for exercise and diet (topic 18, n=5180), devices related to hearing disorders (topic 25, n=3309), and remote learning (topic 22, n=3586, **Supplemental Table 3.4**).

GPT-4 extracted information from 6,940 cardiology department notes containing the term "Apple Watch" using a zero-shot prompt (**Supplemental Table 3.5**) demonstrated that 53.2% of use was initiated by patients, 14.7% by physicians, and 32.1% unknown (**Supplemental Table 3.6**). The majority of Apple Watch usage (54.5%) was for ongoing monitoring purposes. In 26% of notes, data from the Apple Watch led to the current visit, 15.2% of notes contained suggestions for future use, and the remaining 4.1% were categorized as "Other" (**Supplemental Table 3.6**). Regarding downstream clinical care impacts, the most frequent action was to provide new patient instructions (36.8%) or to order additional procedures (21.2%), or lab tests (12.4%). Medication changes were identified in 15.6% of the cases, while new medication orders (4.4%) and discontinuations (5.2%) were less common. New diagnoses (3.0%), referrals (4.3%), and other care changes (9.1%) were also less frequently mentioned (**Supplemental Table 3.7**). Of the 50 notes manually reviewed for accuracy of extracted care data, 64% contained no errors in digital health information extraction. The most common error identified was missing information about changes in clinical care, which occurred in 8/50 (16%) notes (**Supplemental Table 3.8**).

3.4 Discussion

Digital health technologies are increasingly becoming part of routine clinical care, providing patients with greater access to care and allowing clinicians to more effectively manage patient health. Here, we identify and characterize digital health documentation within a large corpus of over 130 million longitudinal clinical notes across a large, academic hospital.

We showed that digital health documentation primarily occurred in clinical notes from slightly younger patients and were more common among patients self-reporting as "White" with "English" as their preferred language. These results align with previous studies highlighting a "digital divide" in the use of digital health tools, particularly for patients with poor language fluency or among elderly patients[17,28–30]. However, future studies may be required to determine whether this discrepancy may be related to differences in demographic information completeness or other clinical factors.

We also demonstrated that digital health documentation has been increasing across terms and departments, at rates faster than the growth of other notes. Several previous studies have also documented differences in digital health adoption across different departments, although these focused on telemedicine or specific digital health tools, not overall digital health usage[31–33]. The growth shown here may be due in part to a shift towards virtual care practices resulting from the COVID-19 pandemic[29,34,35]. Other reasons contributing to the overall growth of digital health documentation may include increasing patient and provider adoption[32,35], advancements in regulatory frameworks and level of evidence for digital health tools[16,20], and improved infrastructure for integration of digital health into clinical workflows[36].

Major themes in digital health documentation uncovered by topic modeling included the use of remote monitoring devices, particularly regulated devices approved for clinical use, such as continuous glucose monitoring[37–39] and atrial fibrillation detection[40]. Manual review of note excerpts also showed that their use in clinical workflows varied significantly in the absence of formal, evidence-driven guidelines. Common clinical impacts of digital health usage, specifically Apple Watch use in cardiology, included patient instructions, procedures ordered, or medication changes. Future studies may clarify the impacts of these and other digital health tools.

The findings of this study should be considered with the following limitations. First, the search terms used in this study may not have captured all relevant digital health references in our notes; more refined search strategies or the development of standardized ontologies may uncover additional digital health use cases in clinical notes or structured medical record data. Second, our analysis is also performed on deidentified data, so inaccurately deidentified digital health tools may be missed. The deidentification process also shifts clinical note dates up to a year for each patient, so trends in usage are only presented as an average across the dataset. Finally, we also specifically excluded notes that contained only telehealth instructions given the prevalence of other studies detailing the effects of telehealth on patient outcomes[34,34,35].

While there is increasing appreciation of the role of social determinants in healthcare[41,42], including digital determinants[30,43,44], resources for patients and physicians to effectively use validated digital health devices and data remain limited[45]. This study provides a starting point for better understanding the contexts of digital health documentation and identifies cases where digital health is actively being used for patient care. As the adoption of digital devices and software continues to grow in both clinical and patient-driven settings, development of new evidence-based guidelines may help standardize workflows for how patients and physicians can use digital health resources most effectively.

3.5 Methods

3.5.1 Curation of digital health terms

To curate a list of digital health terms, we searched PubMed Central and ClinicalTrials.gov for any articles or trials containing the term "digital health" from the last 10 years (2012-2022, **Supplemental Table 3.1**). Search fields were limited to article key words, titles, and abstracts for PubMed. We selected one, two, and three word phrases that occurred in at least 500 abstracts or 50 clinical trial descriptions, and manually inspected the terms to identify synonyms of "digital health" and specific hardware or software interventions. Phrases were excluded if not a noun or were not relevant to digital health, or were captured by other terms on the list. We also screened terms from previously curated lists of connected clinical digital health devices[22] and digital therapeutics[46]. Terms that were selected from this initial screening were further filtered using the UCSF EMERSE search system. EMERSE is an elastic search algorithm built by the University of Michigan that identifies matches or near matches of all terms searched[47]. Terms that did not result in any matches through an EMERSE search, or appeared in contexts that were not digital health related, were excluded. Synonyms of digital health terms that appeared in reviews of notes extracted using EMERSE searches were also included in the final term list. Regular expressions were constructed for the final term list and used to identify notes that contained digital health concepts. The full list of phrases screened and final values selected can be found in the supplemental materials and tables online at https://github.com/BMiao10/DigitalHealthNotes.

3.5.2 Identification of digital health notes from Information Commons at UCSF

Information Commons is a dataset of deidentified, longitudinal medical record data from over 6 million patients at UCSF and contains both structured clinical data and paired clinical notes[23].

Deidentified clinical notes from Information Commons were queried for the presence of any of the digital health search terms selected (**Figure 3.1**) using regular expressions of curated digital health terms. Due to an increase in clinical notes mentioning telehealth visits and related digital health infrastructure following the onset of the COVID-19 pandemic, notes and messages provided to patients outside the clinical visit were excluded based on the note type. These included notes regarding pre-procedure preparation, instructions to set up video calls or telehealth visits, and notes marked as documentation only. All duplicate notes were also removed. Unique patient or encounter IDs were used to identify patient and hospital metadata corresponding to each note. Specifically, baseline patient characteristics (gender, age, race, and patient portal usage) and encounter information (provider and department specialty) were selected for each note. All data used were deidentified and exempt from review based on guidelines from the UCSF IRB.

3.5.3 Distribution and growth of digital health terms in clinical notes

Analysis of the distribution of digital health terms within each department was limited to the top 5 departments with the greatest number of digital health clinical notes. The proportion of each of the top 10 terms within each department was calculated and differences were analyzed by chi-square test. Top terms were selected as the most frequently occurring term across all notes. For notes containing multiple different terms, each unique term was counted as a separate occurrence. However, multiple instances of the same term within a note were only counted as a single occurrence of that term.

The number of notes occurring each year was also analyzed, stratified by digital health terms and by department. Compound annual growth rates (CAGRs) were calculated for occurrences of the top ten digital health terms from 2012-2022 and compared to the growth of the total number of notes and number of digital health notes at UCSF within the same timeframe. The CAGR represents an average, cumulative rate of growth between the time period specified[22]. The total number of digital health notes in each year were also analyzed. Analogous calculations were performed for digital health notes in each department.

3.5.4 Unsupervised classification of digital health notes using LDA topic modeling

An unsupervised approach was taken to identify the context in which these digital health terms were discussed. Notes were tokenized and only sentences containing digital health terms, extracted using the NLTK package[55], were selected for each note. These truncated digital health notes were further preprocessed to remove any uppercase letters and special characters, including asterisks that denote words redacted in the deidentification process. Common English stopwords from the NLTK library, words occurring fewer than 5 times, and words occurring in more than 50% of notes were removed. Topic modeling was performed using Latent Dirichlet Allocation[56] (LDA) and visualized using the pyLDAvis package[57]. The number of topics was selected using a grid search of values between 10-50, and the best value was selected as having the highest Normalized Pointwise Mutual Information (NPMI) coherence value, which describes how closely text in a cluster are related. Digital health notes were assigned to each resulting category based on highest probability to quantify the prevalence of each topic.

3.5.5 Quantifying downstream clinical effects of digital health product usage using large language models

To analyze the clinical impact of specific digital health products, the GPT-4-128-turbo ("GPT-4") large language model was used to extract digital health information using a zero-shot prompt. The prompt, provided in supplemental data, was used to identify whether the digital health usage was initiated by the patient or physician, extract the clinical context for digital health use ("suggested for future use", "data led to current visit", "ongoing monitoring", or "other”), and if the usage resulted in any change(s) to downstream care ("medication change", "medication order", "medication discontinuation", "procedure ordered", "lab test ordered", "new diagnosis", "new referral", "new patient instructions", and/or "other"). The prompt was specifically applied to extract values only from “Apple Watch” clinical notes written in the cardiology department. A subset of 50 notes were manually reviewed for accuracy for each variable extracted. All GPT-4 usage was performed using a HIPAA-compliant endpoint hosted on Microsoft Azure OpenAI Studio.

3.5.6 Statistics

Descriptive statistics for normally distributed continuous distributions are reported as means and standard deviations. Chi-square tests were used to compare categorical proportions, and comparisons between continuous variables were performed using student t-tests. Statistical testing was performed using the SciPy package[58], with $p<0.05$ considered significant.

3.6 Figures

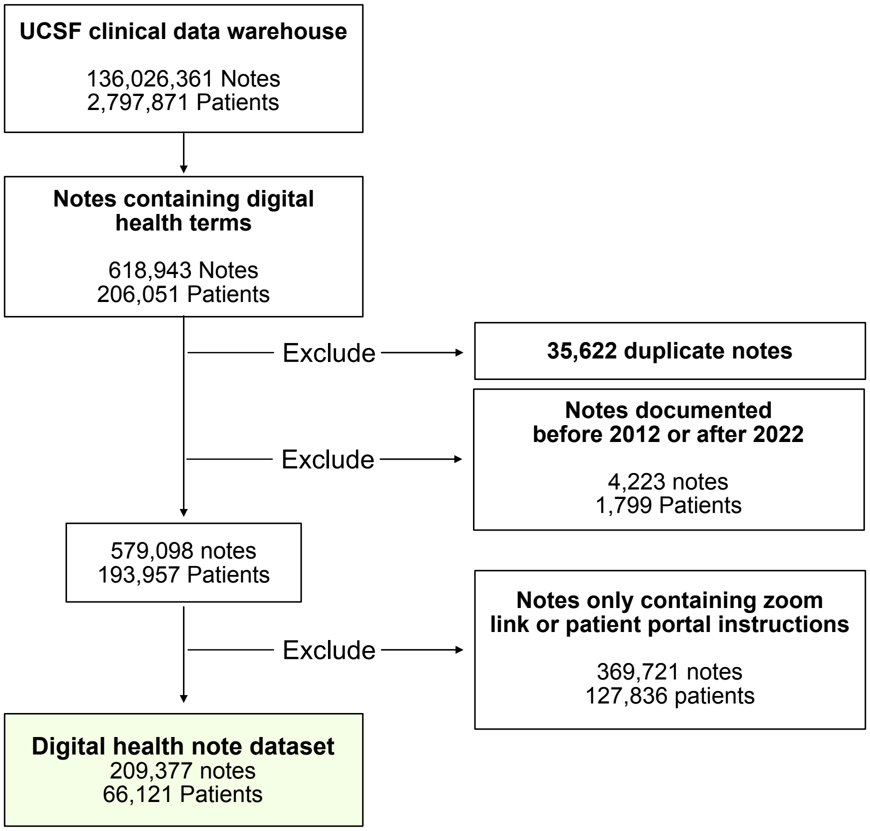


Figure 3.1 *Overview of digital health clinical note dataset selection.*

Clinical notes written between 2012-2022 and containing digital health terms were selected from Information Commons, which contains all deidentified clinical data at UCSF. Notes related to telehealth visits, non-visit messages, and duplicate notes were excluded, leaving a final dataset of 209,377 notes across 66,121 patients.

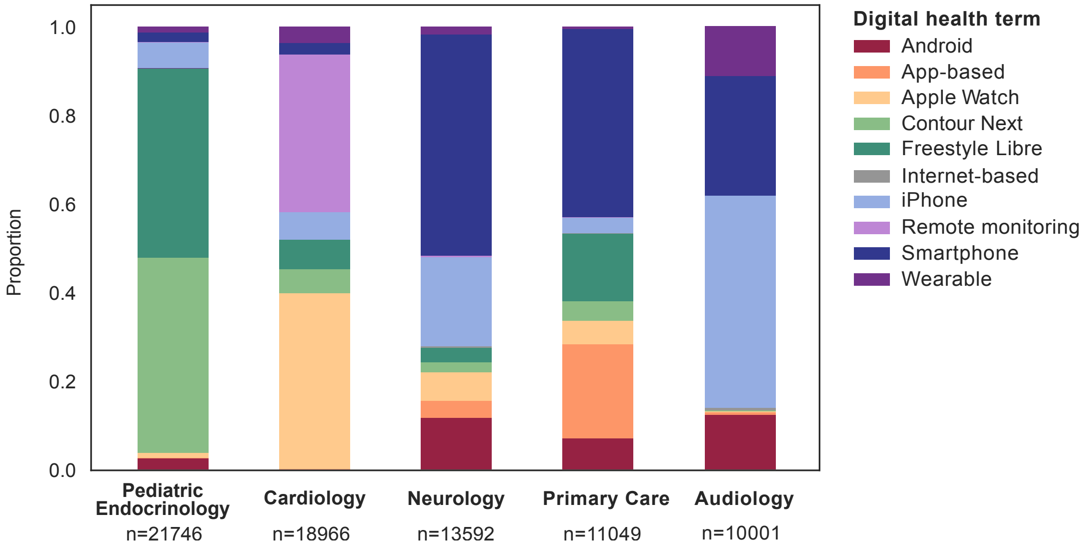


Figure 3.2 *Digital health term distribution across departments.*

Distribution of 10 most frequently occurring digital health terms across the top 5 departments with the greatest number of digital health notes.

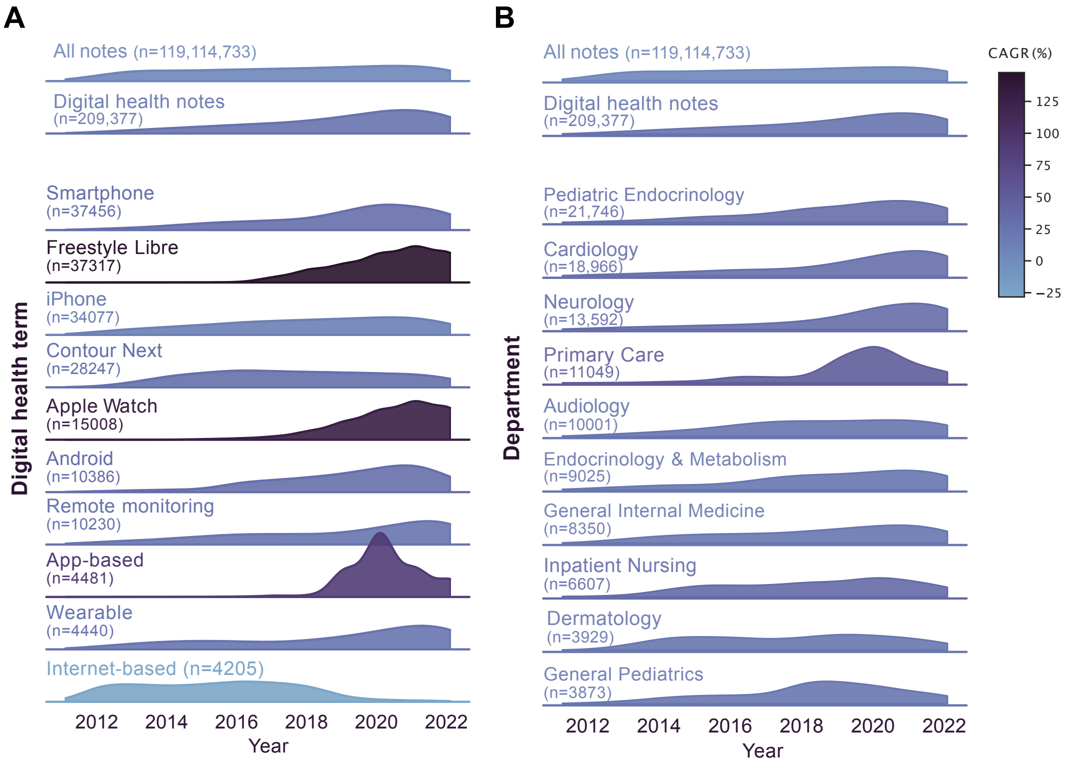


Figure 3.3 *Growth of digital health documentation over time.*

Relative distribution of the top 10 most prevalent A) digital health terms and B) departments with digital health notes from 2012-2022. The number of all notes at UCSF and the number of digital health notes over the same time period are also shown. Plots are colored by CAGR values, representing the average, cumulative rate of growth from 2012-2022.

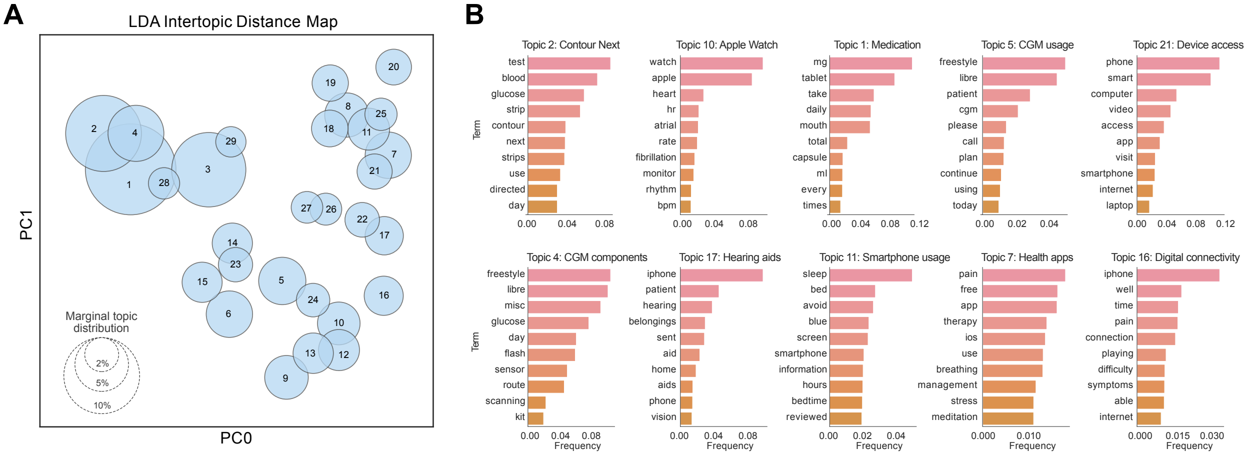
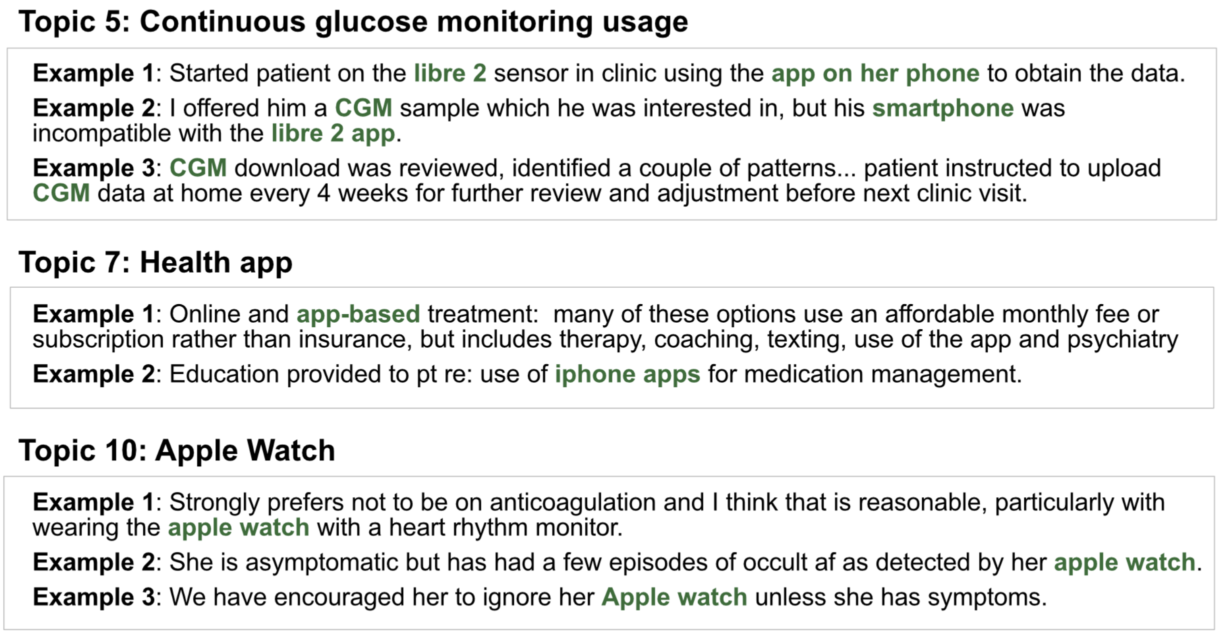


Figure 3.4 *Topic analysis to uncover context of digital health documentation.*

LDA topic modeling was used to identify major clusters of digital health documentation. A) Topic clusters showing the similarity and relative contribution of clusters to the LDA model. B) Top 10 terms for each of the top 10 most prevalent topics in the digital health dataset.



Topic 5: Continuous glucose monitoring usage

Example 1: Started patient on the **libre 2** sensor in clinic using the **app on her phone** to obtain the data.

Example 2: I offered him a **CGM** sample which he was interested in, but his **smartphone** was incompatible with the **libre 2 app**.

Example 3: **CGM** download was reviewed, identified a couple of patterns... patient instructed to upload **CGM** data at home every 4 weeks for further review and adjustment before next clinic visit.

Topic 7: Health app

Example 1: Online and **app-based** treatment: many of these options use an affordable monthly fee or subscription rather than insurance, but includes therapy, coaching, texting, use of the app and psychiatry

Example 2: Education provided to pt re: use of **iphone apps** for medication management.

Topic 10: Apple Watch

Example 1: Strongly prefers not to be on anticoagulation and I think that is reasonable, particularly with wearing the **apple watch** with a heart rhythm monitor.

Example 2: She is asymptomatic but has had a few episodes of occult af as detected by her **apple watch**.

Example 3: We have encouraged her to ignore her **Apple watch** unless she has symptoms.

Figure 3.5 *Examples of sentences containing digital health terms.*

Excerpts of digital health clinical notes from three prevalent topic clusters. Representative examples are chosen to demonstrate the diversity of scenarios documented in digital health clinical notes. Digital health terms are highlighted in green.

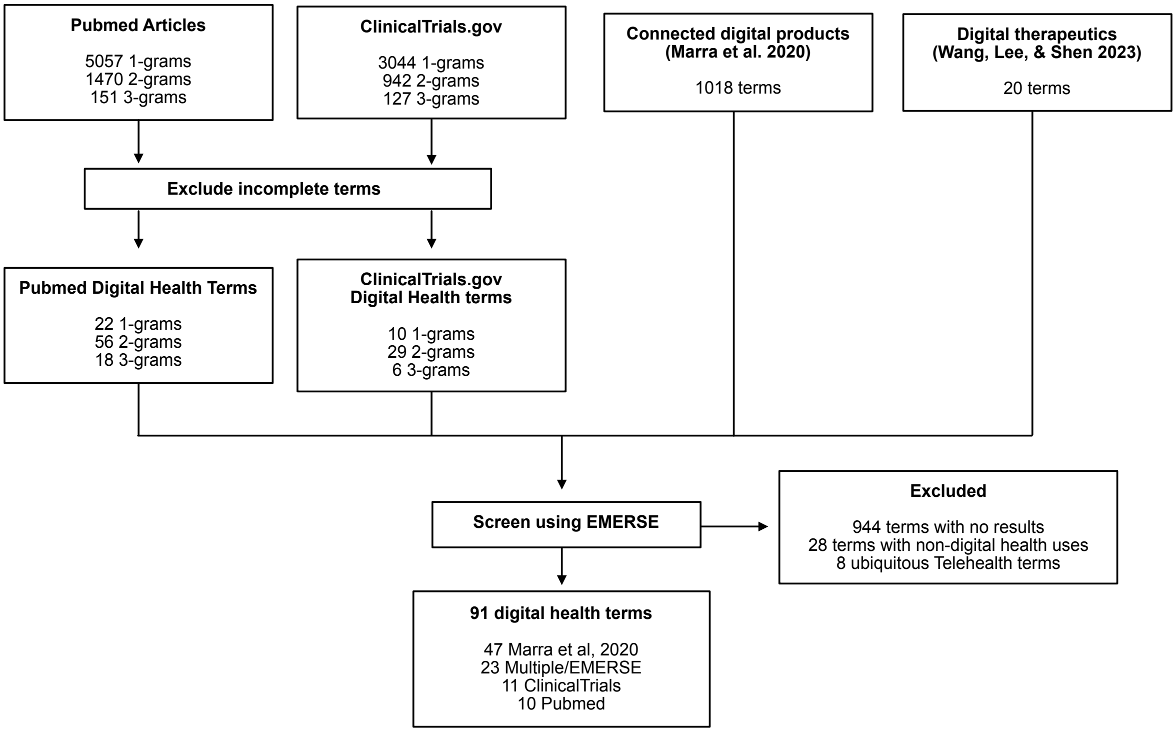
3.7 Tables

Table 3.1 *Demographics of patients with and without digital health notes*

	Patients without digital health note (n=2,169,595)	**Patients with >=1 digital health note (n=66,121)**	**Significance**
Mean age (SD)	44.1 years (25.8)	37.3 years (24.9)	**p<0.0001 (T-test)**
Gender (%)			**p<0.0001 (Chi-square)**
Female	1,189,240 (54.8%)	36,315 (54.9%)	
Male	976,837 (45.0%)	29,744 (45.0%)	
Unknown	3,137 (0.1%)	38 (0.1%)	
Nonbinary	293 (0.0%)	24 (0.0%)	
Race/Ethnicity (%)			**p<0.0001 (Chi-square)**
White	811,781 (37.4%)	31,371 (47.4%)	
Latinx	326,713 (15.1%)	10,650 (16.1%)	
Asian	222,925 (10.3%)	9,618 (14.5%)	
Black or African American	117,790 (5.4%)	5,157 (7.8%)	
Other	149,821 (6.9%)	3,594 (5.4%)	
Unknown/Declined	473,673 (21.8%)	2,774 (4.2%)	
Multi-Race/Ethnicity	22,752 (1.0%)	1,663 (2.5%)	
Southwest Asian and North African	7,307 (0.3%)	608 (0.9%)	

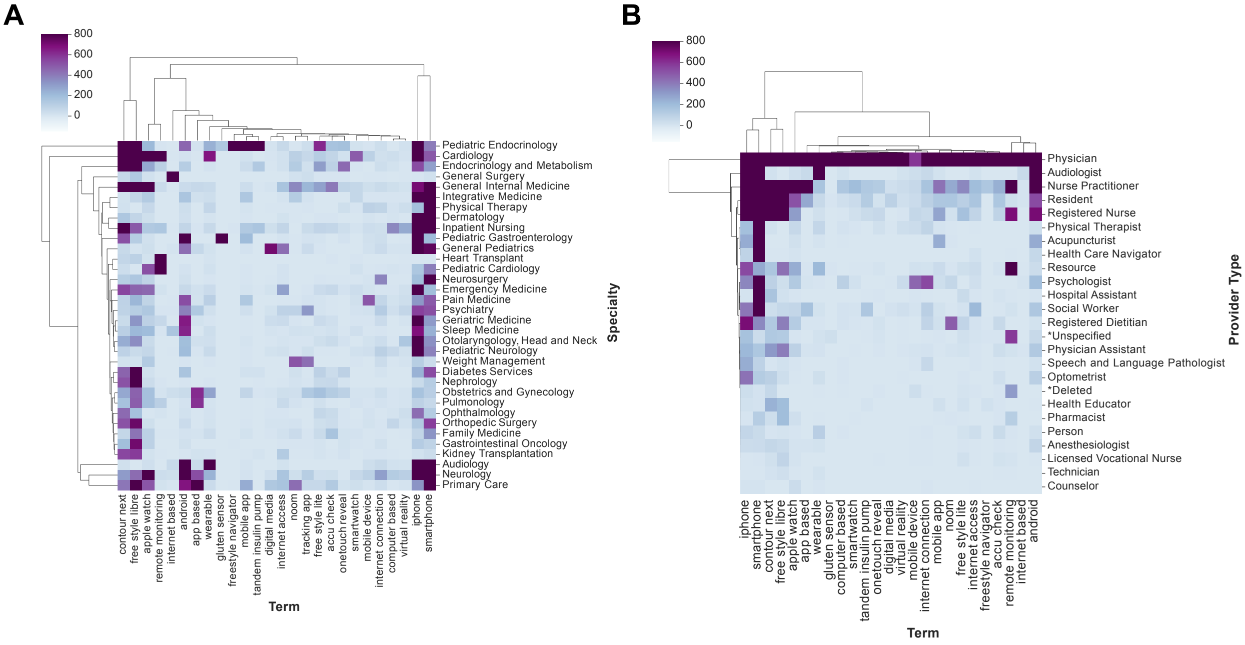
	Patients without digital health note (n=2,169,595)	**Patients with >=1 digital health note (n=66,121)**	**Significance**
Native Hawaiian or Other Pacific Islander	30,882 (1.4%)	441 (0.7%)	
Native American or Alaska Native	5,950 (0.3%)	244 (0.4%)	
Preferred Language (%)			**p<0.0001 (Chi-square)**
English	1,800,100 (83.0%)	60,465 (91.4%)	
Spanish	144,605 (6.7%)	2,932 (4.4%)	
Chinese (Cantonese)	25,077 (1.2%)	732 (1.1%)	
Russian	8,881 (0.4%)	332 (0.5%)	
Chinese (Mandarin)	9,916 (0.5%)	331 (0.5%)	
Vietnamese	5,525 (0.3%)	219 (0.3%)	
Arabic	4,166 (0.2%)	142 (0.2%)	
Korean	2,011 (0.1%)	111 (0.2%)	
Sign Language	820 (0.0%)	85 (0.1%)	
Tagalog	3,225 (0.1%)	79 (0.1%)	

3.8 Supplemental Figures and Tables



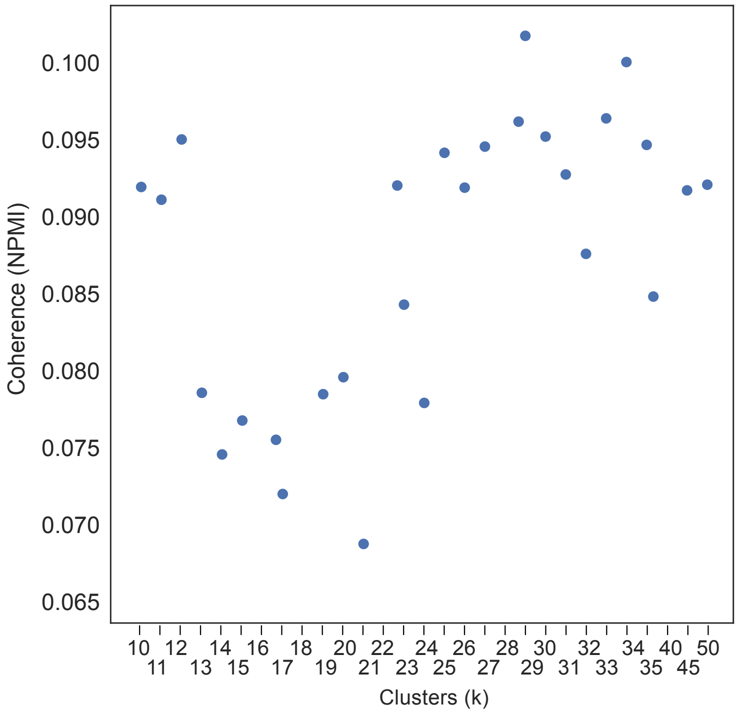
Supplemental Figure 3.1 *Digital health term selection criteria.*

Digital health terms (n=91) were selected using a comprehensive search criteria across PubMed, ClinicalTrials.gov, and reviews of connected digital health products and digital therapeutics.



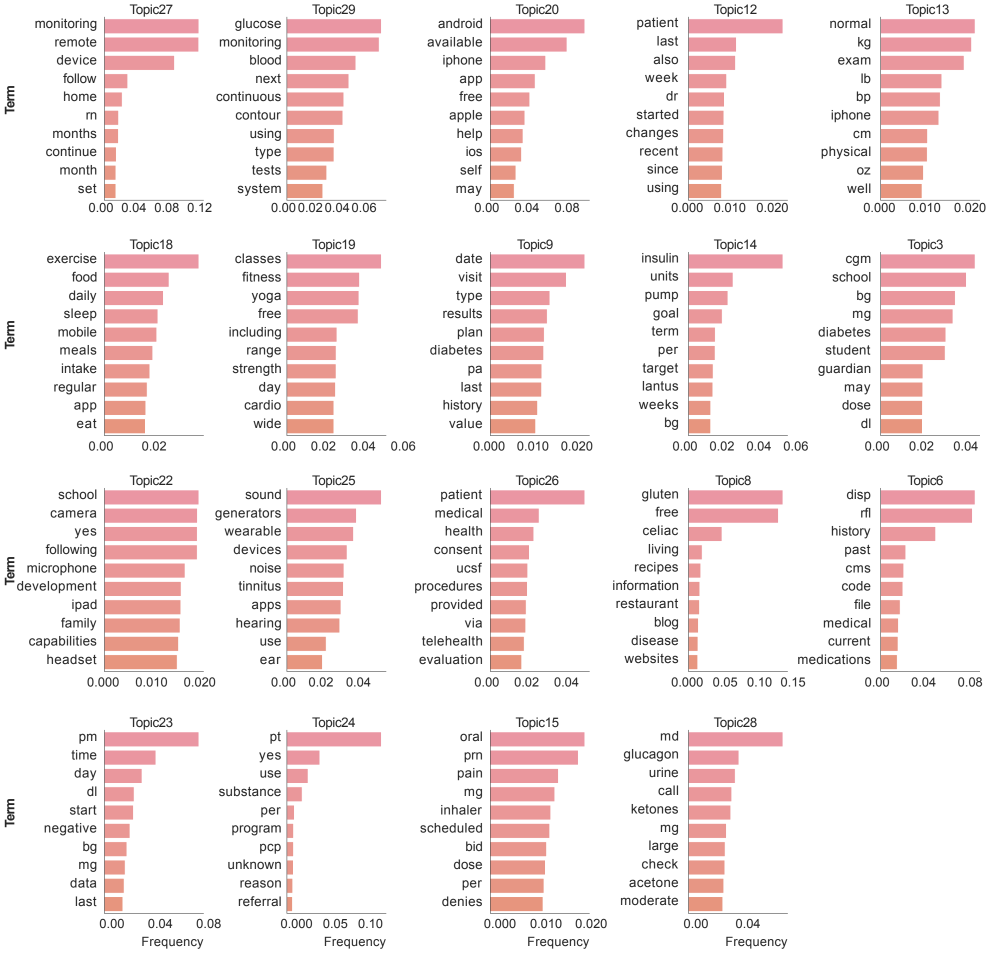
***Supplemental Figure 3.2** Digital health terms by department.*

Prevalence of all digital health clinical notes A) across all departments and B) by provider type. Colorbar scales are set to a maximum of 800 clinical notes to show departments and provider types with fewer clinical notes.



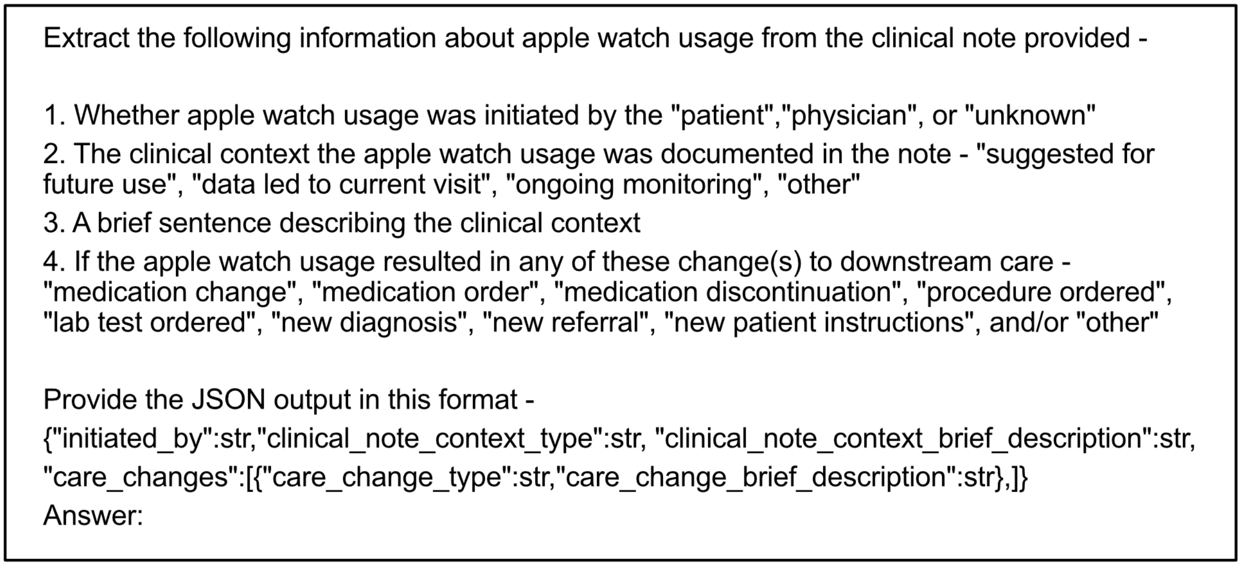
Supplemental Figure 3.3 *Topic modeling coherence scores.*

NMPI coherence scores to identify optimal number of clusters for LDA modeling. The higher the score, the more coherent the topic clusters.



***Supplemental Figure 3.4** Topic modeling results.*

Top 10 terms and frequencies for less prevalent topics identified by LDA topic modeling.



Extract the following information about apple watch usage from the clinical note provided -

1. Whether apple watch usage was initiated by the "patient","physician", or "unknown"
2. The clinical context the apple watch usage was documented in the note - "suggested for future use", "data led to current visit", "ongoing monitoring", "other"
3. A brief sentence describing the clinical context
4. If the apple watch usage resulted in any of these change(s) to downstream care - "medication change", "medication order", "medication discontinuation", "procedure ordered", "lab test ordered", "new diagnosis", "new referral", "new patient instructions", and/or "other"

Provide the JSON output in this format -
{"initiated_by":str,"clinical_note_context_type":str, "clinical_note_context_brief_description":str, "care_changes":[{"care_change_type":str,"care_change_brief_description":str},]}
Answer:

Supplemental Figure 3.5 *Prompt used to extract digital health information.*

References

1. Silberman, J. *et al*. Rigorous and rapid evidence assessment in digital health with the evidence DEFINED framework. *NPJ Digit Med*. 6(1):101 (2023).

2. Witt, D. R., Kellogg, R. A., Snyder, M. P., & Dunn, J. Windows into human health through wearables data analytics. *Current opinion in biomedical engineering*. 9:28-46 (2019).

3. Chen, C., Ding, S., & Wang, J. Digital health for aging populations. *Nat. Med.* **29**, 1623–1630 (2023).

4. Schalkamp, A. K., Peall, K. J., Harrison, N. A., & Sandor, C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. *Nat. Med.* 1–9 (2023).

5. Tison GH *et al.* Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. *JAMA Cardiol.* **3**, 409–416 (2018).

6. Wong, J. C. *et al.* A Pilot Study of Use of a Software Platform for the Collection, Integration, and Visualization of Diabetes Device Data by Health Care Providers in a Multidisciplinary Pediatric Setting. *Diabetes Technol. Ther.* **20**, 806–816 (2018).

7. Kompala, T. & Neinstein, A. B. Telehealth in type 1 diabetes. *Curr. Opin. Endocrinol. Diabetes Obes.* **28**, 21 (2021).

8. West, H. J., Bange, E. & Chino, F. Telemedicine as patient-centred oncology care: will we embrace or resist disruption? *Nat. Rev. Clin. Oncol.* 1–2 (2023).

9. Tang, W. *et al.* The Impact of Telemedicine on Rheumatology Care. *Front. Med.* **9**, 876835 (2022).

10. Singh, B. *et al.* Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *Npj Digit. Med.* **6**, 1–10 (2023).

11. Hartl, D. *et al.* Translational precision medicine: an industry perspective. *J. Transl. Med.* **19**,

245 (2021).

12. Shin, H. J., Han, K., Ryu, L. & Kim, E.-K. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *Npj Digit. Med.* **6**, 1–8 (2023).
13. Solomon, D. H. & Rudin, R. S. Digital health technologies: opportunities and challenges in rheumatology. *Nat. Rev. Rheumatol.* **16**, 525–535 (2020).
14. Huang, J. *et al.* Associations between adoption of eHealth management module and optimal control of HbA1c in diabetes patients. *Npj Digit. Med.* **6**, 1–9 (2023).
15. Johnson, S. A. *et al.* Wearable device and smartphone data quantify ALS progression and may provide novel outcome measures. *Npj Digit. Med.* **6**, 1–10 (2023).
16. Diao, J. A. & Kvedar, J. Mobile health technology for diverse populations: challenges and opportunities. *Npj Digit. Med.* **4**, 1–2 (2021).
17. Nouri, S., Khoong, E. C., Lyles, C. R. & Karliner, L. Addressing Equity in Telemedicine for Chronic Disease Management During the Covid-19 Pandemic. *Catal. Non-Issue Content* **1**, (2020).
18. Katz, A. J. *et al.* Evaluation of Telemedicine Use Among US Patients With Newly Diagnosed Cancer by Socioeconomic Status. *JAMA Oncol.* **8**, 161–163 (2022).
19. Rodriguez, J. A., Saadi, A., Schwamm, L. H., Bates, D. W. & Samal, L. Disparities In Telehealth Use Among California Patients With Limited English Proficiency. *Health Aff. Proj. Hope* **40**, 487–495 (2021).
20. Miao, B. Y., Arneson, D., Wang, M. & Butte, A. J. Open challenges in developing digital therapeutics in the United States. *PLOS Digit. Health* **1**, e0000008 (2022).
21. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).

22. Marra, C., Chen, J. L., Coravos, A. & Stern, A. D. Quantifying the use of connected digital products in clinical research. *Npj Digit. Med.* **3**, 1–5 (2020).

23. Newman-Griffis, D. R., Hurwitz, M. B., McKernan, G. P., Houtrow, A. J. & Dicianno, B. E. A roadmap to reduce information inequities in disability with digital health and natural language processing. *PLOS Digit. Health* **1**, e0000135 (2022).

24. Radhakrishnan, L. *et al.* A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open* **6**, ooad045 (2023).

25. Lybarger, K. *et al.* Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J. Am. Med. Inform. Assoc.* **30**, 1389–1397 (2023).

26. University of California, San Francisco, Academic Research Systems. UCSF DeID CDW-OMOP. University of California, San Francisco. (2022).

27. Doyle-Delgado, K. & Chamberlain, J. J. Use of Diabetes-Related Applications and Digital Health Tools by People With Diabetes and Their Health Care Providers. *Clin. Diabetes* **38**, 449–461 (2020).

28. Harrison, B. & Brown, D. Accuracy of a blood glucose monitoring system that recognizes insufficient sample blood volume and allows application of more blood to the same test strip. *Expert Rev. Med. Devices* **17**, 75–82 (2020).

29. Hincapié, M. A. *et al.* Implementation and Usefulness of Telemedicine During the COVID-19 Pandemic: A Scoping Review. *J. Prim. Care Community Health* **11**, 2150132720980612 (2020).

30. Mann, D. M., Chen, J., Chunara, R., Testa, P. A. & Nov, O. COVID-19 transforms health care through telemedicine: Evidence from the field. *JAMIA.* **27**, 1132–1135 (2020).

31. Gunasekeran, D. V., Tseng, R. M. W. W., Tham, Y.-C. & Wong, T. Y. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. *Npj Digit. Med.* **4**, 1–6 (2021).

32. Bernstein, S. A., Huckenpahler, A. L., Nicol, G. E. & Gold, J. A. Comparison of Electronic Health Record Messages to Mental Health Care Professionals Before vs After COVID-19 Pandemic. *JAMA Netw. Open* **6**, e2325202 (2023).

33. Chunara, R. *et al.* Telemedicine and healthcare disparities: a cohort study in a large healthcare system in New York City during COVID-19. *J. Am. Med. Inform. Assoc.* **28**, 33–41 (2021).

34. Saeed, S. A. & Masters, R. M. Disparities in Health Care and the Digital Divide. *Curr. Psychiatry Rep.* **23**, 61 (2021).

35. Ftouni, R., AlJardali, B., Hamdanieh, M., Ftouni, L. & Salem, N. Challenges of Telemedicine during the COVID-19 pandemic: a systematic review. *BMC Med. Inform. Decis. Mak.* **22**, 207 (2022).

36. Calton, B., Abedini, N. & Fratkin, M. Telemedicine in the Time of Coronavirus. *J. Pain Symptom Manage.* **60**, e12–e14 (2020).

37. Abernethy, A. *et al.* The Promise of Digital Health: Then, Now, and the Future. *NAM Perspect.* **2022**, 10.31478/202206e.

38. Didyuk, O., Econom, N., Guardia, A., Livingston, K. & Klueh, U. Continuous Glucose Monitoring Devices: Past, Present, and Future Focus on the History and Evolution of Technological Innovation. *J. Diabetes Sci. Technol.* **15**, 676–683 (2021).

39. Emerging technologies for the management of type 2 diabetes mellitus - Shah - 2021 - Journal of Diabetes - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1111/1753-0407.13188.

40. Kompala, T. & Neinstein, A. B. Smart Insulin Pens: Advancing Digital Transformation and a Connected Diabetes Care Ecosystem. *J. Diabetes Sci. Technol.* **16**, 596–604 (2021).

41. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

42. Feldman, K. *et al.* Will Apple devices' passive atrial fibrillation detection prevent strokes? Estimating the proportion of high-risk actionable patients with real-world user data. *J. Am. Med. Inform. Assoc.* **29**, 1040–1049 (2022).

43. Marcus, G. M. The Apple Watch can detect atrial fibrillation: so what now? *Nat. Rev. Cardiol.* **17**, 135–136 (2020).

44. Perez, M. V. *et al.* Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **381**, 1909–1917 (2019).

45. Doobay-Persaud, A. *et al.* Teaching the Social Determinants of Health in Undergraduate Medical Education: a Scoping Review. *J. Gen. Intern. Med.* **34**, 720–730 (2019).

46. Maani, N. & Galea, S. The Role of Physicians in Addressing Social Determinants of Health. *JAMA* **323**, 1551–1552 (2020).

47. Aungst, T.D. & Patel, R. Integrating Digital Health into the Curriculum—Considerations on the Current Landscape and Future Developments. *J. Med. Educ. Curric. Dev.* **7**, 2382120519901275 (2020).

48. Chen, M., Tan, X. & Padman, R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J. Am. Med. Inform. Assoc.* **27**, 1764–1773 (2020).

49. Sieck, C. J. *et al.* Digital inclusion as a social determinant of health. *Npj Digit. Med.* **4**, 1–3 (2021).

50. Sushil, M. *et al.* CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 28;1(4):AIdbp2300110. (2024).

51. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).

52. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

53. Wang, C., Lee, C. & Shin, H. Digital therapeutics from bench to bedside. *Npj Digit. Med.* **6**, 1–10 (2023).

54. Hanauer, D. A., Mei, Q., Law, J., Khanna, R. & Zheng, K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J. Biomed. Inform.* **55**, 290–300 (2015).

55. Loper, E. & Bird, S. NLTK: The Natural Language Toolkit. *ACL.* (2004).

56. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation.

57. Sievert, C. & Shirley, K. LDAvis: A method for visualizing and interpreting topics. in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* 63–70 (ACL, 2014). doi:10.3115/v1/W14-3110.

58. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

Chapter 4

Quantifying Clinical-Decision Making Using Large Language Models

4.1 Abstract

Understanding the factors that drive treatment selection and switching is of significant medical interest. However, many factors related to medication switching are often only captured in unstructured clinical notes and can be difficult to extract. We evaluate the zero-shot abilities of a large language model, GPT-4 (via HIPAA-compliant Microsoft Azure API), to identify reasons for switching between classes of contraceptives from clinical notes of 1,964 contraceptive switches in 1,515 patients in the UCSF Information Commons. When evaluated by clinical experts, GPT-4 extracted switching reasons with an accuracy of 91.4% and 2.2% hallucination rate. Using extracted reasons, we identified patient preference, adverse events, and insurance as key reasons for switching using unsupervised BERTopic modeling. Notably, we also showed using our approach that "weight gain/mood change" and "insurance coverage" are disproportionately found as reasons for contraceptive switching in underserved demographic populations.

4.2 Introduction

Prescription contraceptives play a critical role in supporting women's reproductive health. With the increasing availability of different contraceptives, providing patients with new options to manage their reproductive health, there is also a growing need to provide patients and providers with data-driven guidelines for informed decision-making[1–3]. Contraceptives may vary by active ingredient, with either progestin-only active ingredients or estrogen-progestin combinations available, as well as by mode of administration, which may be intrauterine (Intrauterine devices, IUDs), oral, transdermal, intravaginal, subdermal, or as an injection[4]. Each of these contraceptives have unique adverse event profiles that may contribute to clinical decision making[5]. In addition, several other factors, including personal preference, cost, availability, comorbidities and clinical constraints, may contribute to a patient's decision to start, stop, or switch contraceptives[6]. With nearly 50 million women in the United States using contraceptives[7], understanding the factors that drive contraceptives selection and switching is of significant interest.

Previous studies of medical claims data have shown that 44% of women starting a contraceptive discontinued its use within one year, although 76% resumed use of the same or another contraceptive within three months[8]. Some studies have begun to move beyond analysis of discontinuation rates and have included interviews[9] or social media data[10] to better understand the complexity of contraceptive switching. Using text mining approaches, these studies have identified specific patient subgroups that switch at different rates[11] or showed that patients have differing preferences and reasons for seeking contraceptives[4]. However, these studies may not capture the breadth or depth of clinical information found in medical record data and require development of custom machine learning models or time-consuming manual analysis to generate insight from the complexity of real-world text data[12–14].

Recently, the development of general large language models (LLMs) has shown significant promise in being able to extract medication information without the need for manually annotated training data ("zero-shot extraction")[12,15,16]. Despite concerns including factually incorrect information, clinicians and researchers remain optimistic that these computational advances can translate to clinically-meaningful use cases[17–20]. Here, we evaluate the ability of GPT-4 to extract contraceptive selection strategies and identify reasons for switching between classes of contraceptives using clinical notes from a large academic medical center.

4.3 Results

4.3.1 Contraceptive switching cohort

We selected a contraceptive patient cohort using the UCSF Information Commons dataset. We identified 37,834 patients with at least 1 medication order for an intrauterine, oral, intravaginal, subdermal, transdermal, or injectable contraceptive. We removed 5,594 patients who did not have any follow up encounters at least 6 months after the last contraceptive order. This left 37,834 patients with 100,593 medication orders. We further filtered out the 11,916 orders without associated clinical notes and 53,125 duplicate medication orders, leaving a contraceptive cohort consisting of 39,790 medication orders across 20,283 unique patients (**Figure 4.1**).

Among this contraceptive cohort, 1,515 (7.6%) patients experienced a total of 1,964 contraceptive switches. Compared to patients who did not have a contraceptive switch, patients with contraceptive switches tended to be younger, with a mean age of 25.9 years (SD: 7.7) compared to 29.1 years (SD: 8.4, p<0.001, **Table 4.1**). There was also a statistically significant difference in the proportion of patients with and without contraceptive switches by patient race/ethnicity (p<0.001). The largest difference occurred in patients with a race/ethnicity listed as "Black or African American," with 19.3% of such patients having a contraceptive switch compared to 8.2% without. Patients identifying as "Latinx" were also more likely to have a contraceptive switch (19.3%) compared to the proportion of "Latinx" patients without contraceptive switches (15.1%). "White" (33.0%) or "Asian" (16.0%) patients had lower rates of contraceptive switching in this cohort compared to the same groups without switches, with 45.1% of patients without contraceptive switches identifying as "White" and 20.3% identifying as "Asian."

Switching differed significantly by the first contraceptive prescribed, with the highest rates of switching following initial prescription of transdermal contraceptives (33.5%) and the lowest rates following initial prescription of intrauterine (5.1%) and oral (6.3%) contraceptives. The most common switch occurred in patients who were on oral contraceptives and switched to intravaginal contraceptives (n=205, n=10.5%). The least common switch occurred from intrauterine to injectable contraceptives (n=6, 0.31%, **Supplemental Table 4.4**). All supplemental tables are made available on Github at https://github.com/BMiao10/contraceptive-switching.

4.3.2 Human evaluation of GPT4 extraction of contraceptive switching

Prompt evaluation was performed on a held out set consisting of notes from 5% of patients (n=93 clinical notes), and evaluated against annotations from a clinical reviewer. There was no significant difference in performance across the six prompts used to extract contraceptive information using zero-shot GPT-4, with micro F1 scores ranging from 0.817 to 0.849 (mean=0.827, SD: 0.012) for extraction of contraceptive started, and 0.827 to 0.881 (mean=0.854, SD: 0.020) for extraction of contraceptive stopped (**Figure 4.2**). The best prompt for medication stopping extraction used the specialist system configuration and default prompt. Reasons extracted by this prompt were also evaluated by a clinical reviewer for both accuracy and rate of hallucination. Human evaluation showed that GPT-4 was capable of extracting these reasons with 91.4% accuracy and without hallucination 97.8% of the time (n=93). Given the high accuracy and minimal hallucination of this prompt for extracting information about contraceptive stopping and reasons for stopping on the development dataset, this prompt was selected to extract contraceptive information from the remaining clinical notes.

4.3.3 GPT-4 contraceptive switching information extraction outperforms baseline models

Zero-shot GPT-4 performance using the best prompt was also compared to baseline models trained on different proportions silver-standard labels derived from structured data. GPT-4 outperformed all baseline models, regardless of the proportion of training data used for baseline models (**Figure 4.3**, with micro F1 scores of 0.828 and 0.439 on contraceptive start and stop extraction, respectively. The next best model was random forest trained on TF-IDF representations, with a 0.714 (SD: 0.024) score on medication start and 0.424 (SD: 0.009) on medication stopping.

Concordance between silver-standard labels and human annotations available showed a Cohen's Kappa coefficient of 0.585 for medication starting labels and 0.217 for contraceptive stopping (n=93). When we removed notes without relevant contraceptives, determined by the human evaluator, concordance between these two methods increased to 0.960 for contraceptives started and 0.644 for contraceptives stopped (**Supplemental Table 4.5**).

4.3.4 Identification of reasons for contraceptive switching

Unsupervised BERTopic topic modeling of extracted reasons for stopping across the full dataset identified 19 topics, which were manually grouped into 10 cohesive topics (**Supplemental Table 4.6**). Excluding the 1136 notes that did not contain a relevant reason (topic 0, **Supplemental Table 4.7**), the most frequently occurring topics contained terms related to spotting and irregular bleeding (topic 1), desire to switch contraceptives (topic 2), and forgetting to take daily pills (topic 3). Topics 4, 6, 7 described other adverse events of contraceptive use, including irritation and rash, weight gain and mood changes, and irregular menses and pain. Topic 5 related to IUD malpositioning and removal, and topic 9 related to implant removal. Finally, topic 8 included terms related to insurance coverage (**Figure 4.4**).

Subset analysis stratified by race/ethnicity identified enrichment of specific topics within certain patient subgroups. Weight gain and mood change (topic 6) were enriched in patients who self-reported as being “Latinx” or “Other” and showed lower enrichment in patients self-reporting as “Black or African American”. Topic 9 (Implant removal) was enriched in patients who self-reported a race/ethnicity of “Asian”, and topic 8 (insurance coverage) was enriched in patients of “Black or African American”, “Latinx”, or “Multi-Race/Ethnicity” race/ethnicity (**Figure 4.4**).

4.4 Discussion

We demonstrated that GPT4 can accurately extract which medications were started and stopped during an encounter from associated clinical notes. GPT-4 performance, evaluated by both automated analysis and gold-standard manual annotation, was stable between six different prompts although more complex prompting methods may further improve medication information extraction[24,33]. We further showed that the majority of reasons for contraceptive switching extracted by GPT-4 were also correct, with minimal hallucinations.

Lastly, we uncovered latent contraceptive-specific reasons for switching medications by clustering embeddings derived from GPT-4 extracted values. Topic clusters ranged from treatment failure to patient preference, as well as adverse events and insurance reasons. In line with previous studies[34], we showed that weight gain and mood changes as reasons for switching were enriched in patient populations who self-reported their race/ethnicity as “Latinx” or “Other”. Additionally, we showed that insurance coverage as a reason for switching disproportionately affected patients identifying as “Latinx” or “Black or African American.” Our results highlight recent concerns regarding financial barriers to contraceptive access and resulting racial inequities in reproductive health[35]. Future validation in independent datasets will be needed to determine whether this difference persists in larger samples or in other cases of medication switching.

There are several limitations to this study. This dataset is limited to values derived from a large, academic medical center, which may introduce bias in the types of patients or contraceptives captured. We assume that clinical notes contain information on all medications ordered at the same or previous encounters, but some medications may not be discussed or documented. This is reflected in poor concordance between structured data labels and human evaluation, particularly for medication stopping values. Additionally, because the de-identification process is not perfect, manual review of some notes identified several medication names that were inappropriately redacted. This was particularly prevalent among contraceptive brand names that resemble common patient names (eg. "Camila" or "Heather") that are deliberately redacted. Finally, another limitation of our work surrounds interpretability of results, which is significant to clinical care. There is little public information provided about GPT4's training data, approach, or model architecture, and we have not yet tested any open-source language models on this task. As a result, we refrain from making conclusions about why LLMs like GPT-4 produces certain results and focus instead on evaluating overall performance and insights that can be derived from extraction of information from clinical notes.

In conclusion, our findings demonstrate that reasons for contraceptive switching are disproportionately found in specific patient demographics. Our approach can be applied towards treatment strategy analysis across or within different classes of medications beyond contraceptive switching to improve understanding of treatment strategy and shape more detailed treatment effect estimation models.

4.5 Methods

4.5.1 Contraceptive switching cohort selection

A contraceptive switching cohort was selected from the UCSF Information Commons dataset[21], which contains deidentified structured data and clinical notes from over 6 million patients between 2012-2023. Clinical text notes were certified as deidentified as previously described[22] and are usable by UCSF researchers as non-human subjects research.

We identified all patients prescribed at least one contraceptive documented in the structured medication data based on a "therapeutic class" label. Non-drug contraceptives (e.g diaphragms/cervical caps, condoms, vaginal pH modulators, and spermicides), progestin and estrogen-containing agents not used for contraceptive purposes, and emergency contraceptives were removed (**Supplemental Table 4.1**). The remaining contraceptives were mapped to the following modalities: Oral, Implant, Intrauterine device (IUD), Injection (intramuscular or subcutaneous), Transdermal, and Intravaginal based on regular expression values (**Supplemental Table 4.2**). Contraceptives prescribed without a start date or associated clinical note and duplicate orders at each encounter date were removed. To filter out short notes without any relevant information, only clinical notes containing more than 50 tokens, created using encodings from OpenAI's open-source tokenizer tiktoken.

The dataset was further filtered to patients with encounters at least 6 months after the prescription of the first contraceptive, ensuring those without a switch weren't lost to follow-up. Prescriptions were sorted by documented start date, and encounters that contained a contraceptive switch were retrieved. A contraceptive switch was defined as a difference in prescribed contraceptive modalities between consecutive encounters.

Self-reported demographic information on race/ethnicity and preferred language were extracted from structured data, which was also used to calculate age at date of first contraceptive prescription. This study was conducted using retrospective, deidentified clinical data and was determined to be exempt from IRB review. All data were stored or processed on HIPAA compliant hardware at UCSF or through a HIPAA compliant Microsoft Azure instance ("UCSF Versa"). No data was transferred or stored by OpenAI; and OpenAI settings were maintained so that no prompt information would be stored, even temporarily. All code along with supplemental data and tables are made available on Github at https://github.com/BMiao10/contraceptive-switching.

4.5.2 Prompt evaluation for extraction of contraceptive selection strategy

Prompting can have significant effects on the accuracy of large language models[23,24]. We tested six prompts (**Supplemental Table 4.3**), varying both system information and output formats, to extract the following information: 1) which contraceptive was stopped, 2) which new contraceptive was started, and 3) why the contraceptive switch occurred. To avoid overfitting, these six prompts were evaluated on a held-out subset of contraceptive switching clinical notes from 5% of the patients. The model used was GPT-4, with temperature set at 0, maximum response length capped at 500 tokens, top_p set to 1, and all other parameters kept as default. A zero-shot approach was used, with no additional information or training data provided outside of the encounter's associated clinical note. Resulting values were mapped to the six contraceptive modalities using regular expression values (**Supplemental Table 4.2**). All GPT-4 queries were performed between November 13-15, 2023.

A clinical evaluator assessed the accuracy of GPT-4 extraction for contraceptives started and stopped within each note. Micro F1 scores, which represent the harmonic mean of precision and recall scores, are reported. The best prompt was selected based on the highest average score attained across all medications started/stopped determined by manual evaluation. The clinical reviewer was also instructed to identify whether the extracted reason was accurate based on the clinical note and whether any hallucination occurred, which was defined as information produced by the language model that could not be derived from the clinical note.

4.5.3 Comparison of GPT-4 contraceptive information extraction to baseline models

The best prompt selected from the development dataset was applied to the remaining 95% "test set" of the contraceptive switching cohort using the same GPT-4 setup. We compared our LLM-based methods against several traditional machine learning techniques, including logistic regression, random forest, and BERT-style models. Since human clinical annotations were not available for this larger dataset, weak labels from structured data, specifically which contraceptives were started and stopped at the associated clinical encounter, were used for training and evaluation in each of these models. Structured data may not reflect the contents of clinical notes if patients are prescribed contraceptives at a different facility or stop dates are not documented, so we compared these silver-standard labels to human annotation for the 93 clinical notes in the prompt evaluation set using Cohen's Kappa coefficient to assess reliability between the two sources.

Two sets of logistic regression and random forest models were developed using either bag-of-words and term-frequency inverse document frequency (TF-IDF)[25] text representations. Multiclass classification was performed, with models predicting the modality of contraceptives started or stopped (oral, IUD, subdermal, intravaginal, injection, transdermal). We performed 5-fold cross validation using a 70/10/20 split between train, validation, and test data. Due to differences in training sizes between baseline models and GPT-4, this split is independent of the previous prompt evaluation and GPT-4 test sets. Hyperparameter tuning was performed using a grid search of varying regularization values (C=[0.01, 0.1, 1, 10, 100, 1000]) for logistic regression and both number of estimators and max depth for random forest (n_estimators=[50, 100, 250, 500], max_depth=[20, 50, 100]).

The UCSF-BERT model[26,27] trained on a large corpus of clinical notes was also used as a baseline. Again, we performed 5-fold cross validation using a 70/10/20 split. Hyperparameter tuning was performed using Optuna[28], and both learning rate and weight decay were varied (learning rate=(1e-5, 5e-5), weight decay=(4e-5, 0.01)). Models were trained for 5 epochs, with early stopping. To accommodate for the 512 maximum token length allowed by UCSF BERT, a sliding window was used with final prediction selected by majority vote across all windows.

To simulate few-shot learning, we trained each of the baseline models on random subsamples of 100%, 50%, 25%, 10%, 5%, and 1% of the training data. Micro-averaged F1 scores are reported for each model on the held-out test set.

4.5.4 Unsupervised clustering of extracted reasons for contraceptive switching

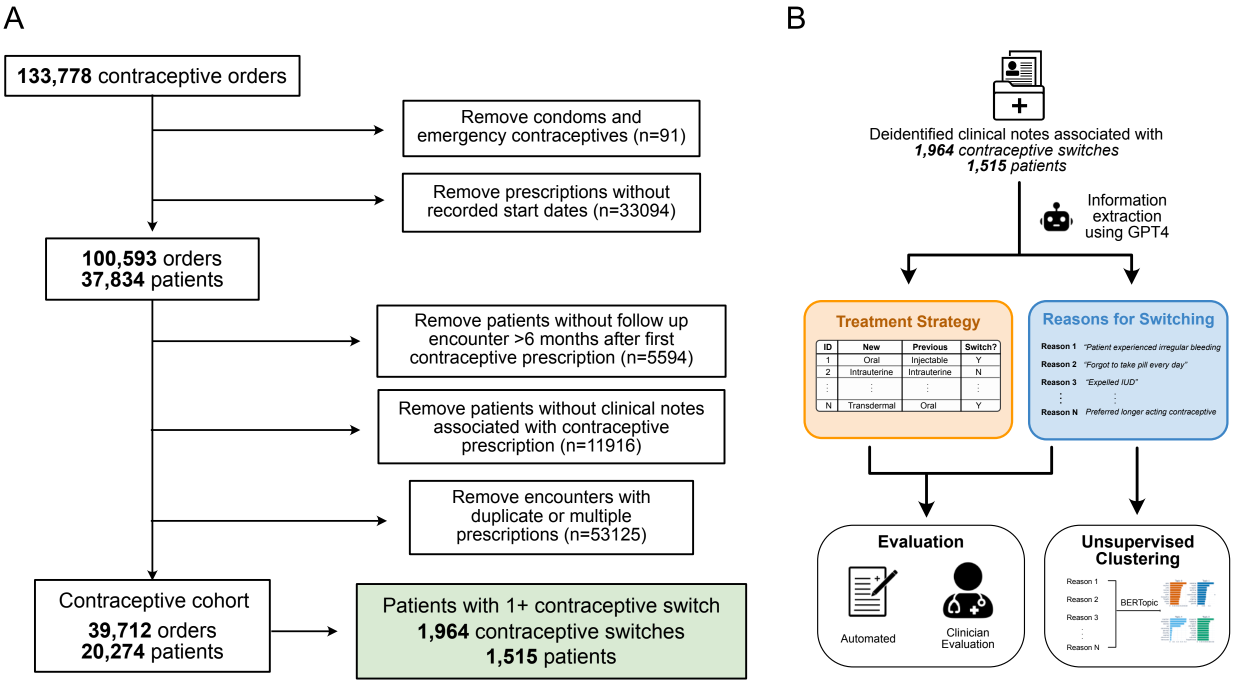
GPT-4 was also used to extract reasons for contraceptive switching from the test set using the best prompt. To identify key reasons for medication switching, we applied BERTopic, a topic modeling method that clusters document embeddings, to all reasons extracted from both the prompt evaluation and test sets. The UCSF-BERT model was used to generate embeddings from the list of extracted reasons and embeddings were clustered by BERTopic[29]. Briefly, dimensionality reduction was applied to the embeddings using Uniform Manifold Approximation and Projection (UMAP), with 5 components and 3 neighbors with Euclidean distance metrics. HDBSCAN[30] was used to cluster reduced embeddings, with number of topics dynamically chosen by the algorithm, and TF-IDF used to identify key terms from each cluster. All other default parameters were used. Topics were manually reviewed and similar topics were grouped together.

Subgroup analysis was performed to understand whether topics were associated with particular patient demographics. Adapting from previous enrichment methods[31], we used topic probabilities assigned to each document by the BERTopic model to calculate a weighted enrichment score that describes the relative contribution of each topic to patient subgroups. Specifically, enrichment scores were calculated as $\theta_{k,j}^{\square} = \frac{q_{n,k} \cdot y_{n,j}}{\sum_{n=1}^{N} \square q_{n,k} * \sum_{n=1}^{N} \square y_{n,j}}$, where q(n,k) describes the weight of each topic k for note n, and y(n,j) are the patient subgroups assigned to each note. The scores were normalized by total topic weight, as well as by number of patients in each subgroup, and reported scores were negative log transformed.

4.5.5 Statistics

We present means and standard deviations for continuous distribution and utilize two-sided t-tests to analyze differences in continuous distributions. To evaluate differences in categorical data, Chi-square tests were applied. Statistical analyses were conducted using the SciPy package[32], and a p-value less than 0.05 was used to indicate statistical significance.

4.6 Figures


A
133,778 contraceptive orders
Remove condoms and emergency contraceptives (n=91)
Remove prescriptions without recorded start dates (n=33094)
100,593 orders
37,834 patients
Remove patients without follow up encounter >6 months after first contraceptive prescription (n=5594)
Remove patients without clinical notes associated with contraceptive prescription (n=11916)
Remove encounters with duplicate or multiple prescriptions (n=53125)
Contraceptive cohort
39,712 orders
20,274 patients
Patients with 1+ contraceptive switch
1,964 contraceptive switches
1,515 patients
B
Deidentified clinical notes associated with
1,964 contraceptive switches
1,515 patients
Information extraction using GPT4
Treatment Strategy
Reasons for Switching
Evaluation
Automated
Clinician Evaluation
Unsupervised Clustering
Reason 1
Reason 2
Reason 3
Reason N
BERTopic

ID	New	Previous	Switch?
1	Oral	Injectable	Y
2	Intrauterine	Intrauterine	N
⋮	⋮	⋮	⋮
N	Transdermal	Oral	Y

Reason 1 *"Patient experienced irregular bleeding*

Reason 2 *"Forgot to take pill every day"*

Reason 3 *"Expelled IUD"*

⋮

Reason N *Preferred longer acting contraceptive*

Figure 4.1 *Study overview*

A) We selected a contraceptive patient cohort from the UCSF Information Commons dataset. Among 20,283 patients with unique contraceptive prescriptions and associated clinical notes, 1,515 (7.6%) patients experienced a total of 1,964 total contraceptive switches. B) Study overview to assess the ability for GPT4 to extract contraceptive switching values from clinical notes, and to identify key reasons for switching using unsupervised clustering methods.

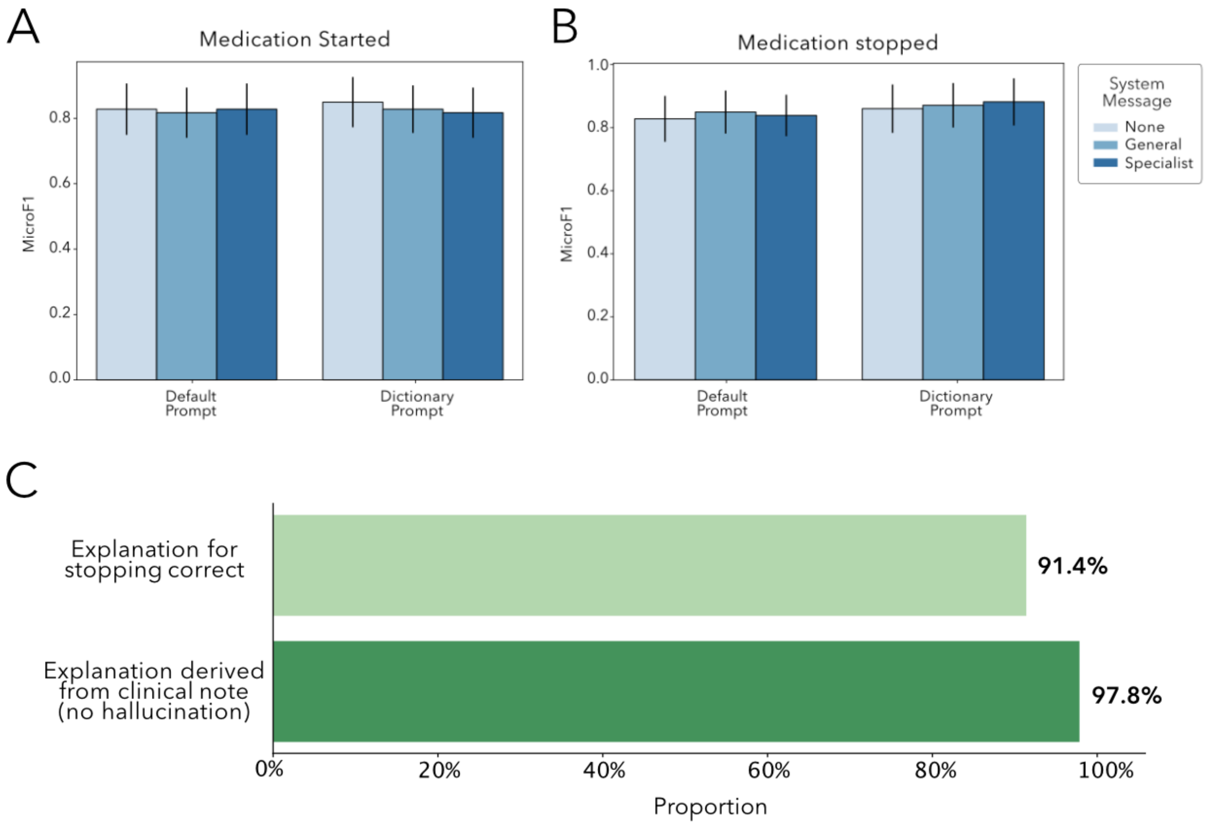


Figure 4.2 *Development of prompt to extract contraceptive switching information.*

GPT4-extracted values for contraceptive class A) started and B) stopped compared to human annotation (n=93). C) Human evaluation was also performed to assess whether GPT-4 extracted reasons for contraceptive switching was accurate and contained only information specifically mentioned in the associated clinical note (not hallucination).

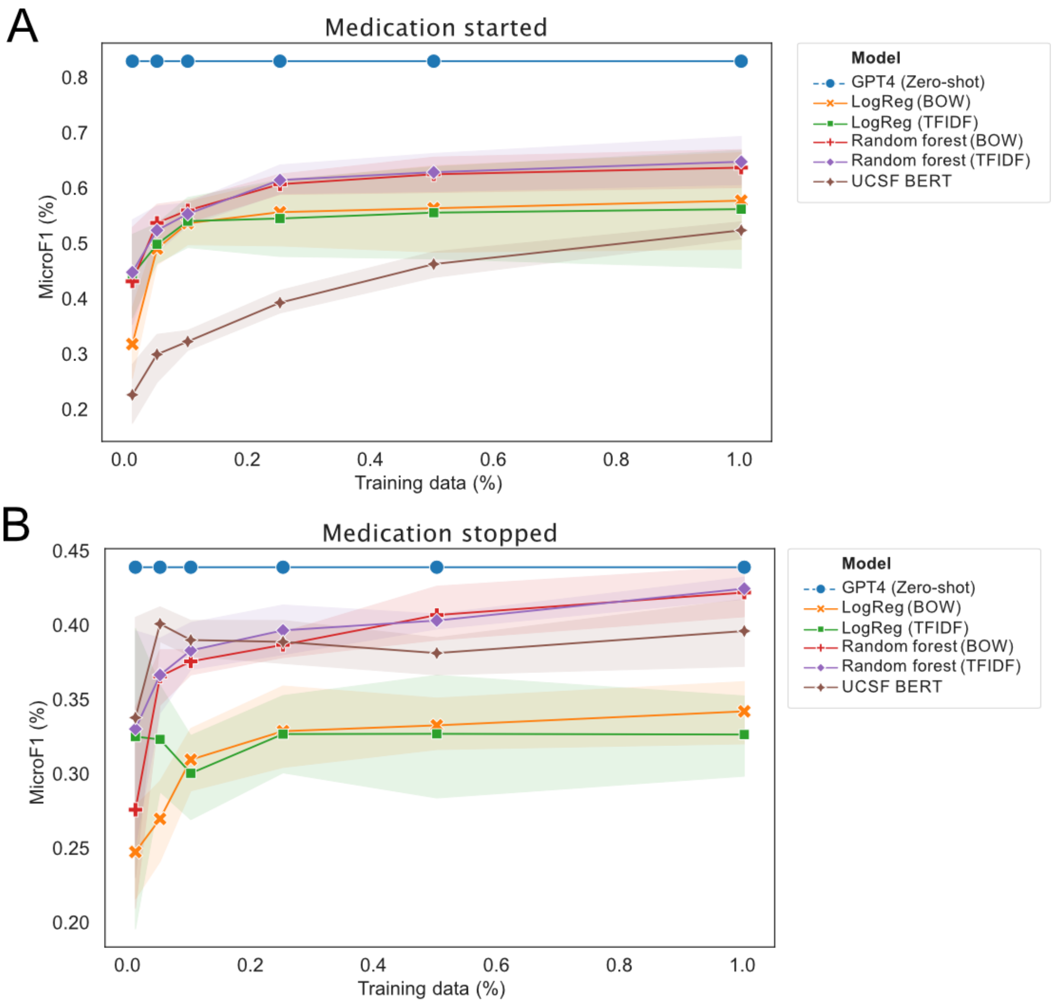


Figure 4.3 *GPT-4 performance compared to baseline.*

Following prompt evaluation, GPT-4 performance on the remaining test set was also compared to baseline model performance for extraction of contraceptive A) started and B) stopped. Silver-standard labels from structured data were used for training and evaluation of baseline models, and for evaluation of zero-shot GPT-4.

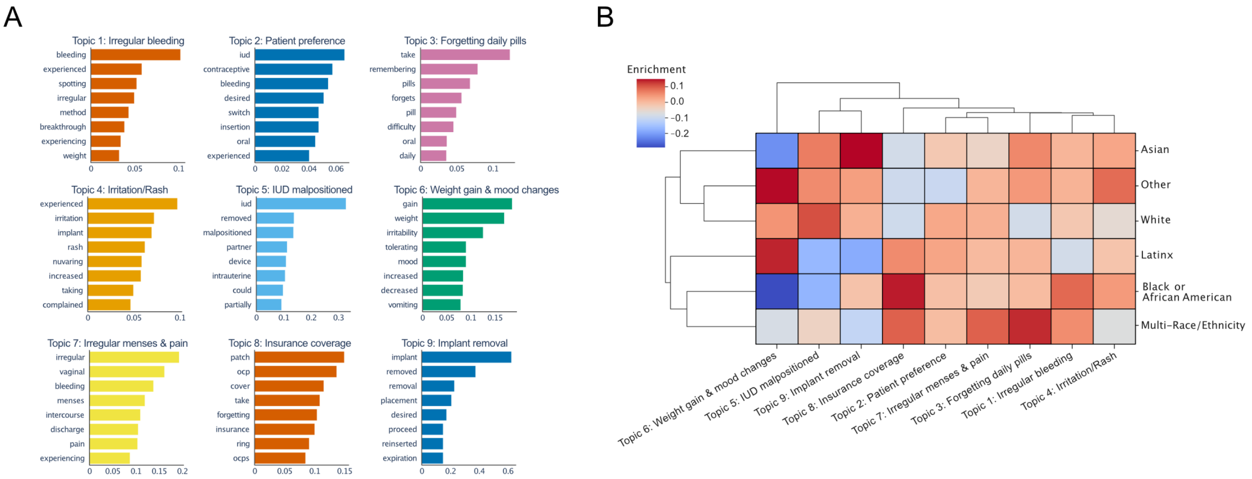


Figure 4.4 *Clustering reasons for contraceptive switching using BERTopic*

A) BERTopic modeling was used to cluster GPT-4 extracted reasons for contraceptive switching, with nine key topics identified. Top terms for each cluster are shown. B) Topics were assessed for enrichment amongst patient subgroups by race/ethnicity. Higher enrichment scores indicate higher prevalence of a topic written in notes within a patient subgroup.

4.7 Tables

Table 4.1 *Contraceptive prescription cohort demographics.*

	Contraceptive switch (n=1,515)	**No switch (n=15,907)**	**Significance** Proportion
Mean age (SD)	25.9 years (7.7)	29.1 years (8.4)	**p<0.001 (Two tailed T-test)**
Race/Ethnicity (%)	Missing (n=32)	Missing (n=815)	**p<0.001 (Chi-square)**
White	490 (33.0%)	6813 (45.1%)	
Latinx	286 (19.3%)	2281 (15.1%)	
Black or African American	286 (19.3%)	1237 (8.2%)	
Asian	237 (16.0%)	3071 (20.3%)	
Other	115 (7.8%)	1224 (8.1%)	
Multi-Race/Ethnicity	69 (4.7%)	466 (3.1%)	
Preferred Language (%)		Missing (n=5)	**p<0.001 (Chi-square)**
English	1474 (97.3%)	15405 (96.9%)	
Spanish	14 (0.9%)	281 (1.8%)	
Other	27 (1.8%)	216 (1.4%)	
First prescribed contraceptive, (%)			**p<0.001 (Chi-square)**
Implant	160 (10.6)	799 (5.0)	20.0%
Injectable	199 (13.1)	853 (5.4)	23.3%
Intrauterine	64 (4.2)	1266 (8.0)	5.1%
Intravaginal	244 (16.1)	1935 (12.2)	12.6%
Oral	661 (43.6)	10496 (66.0)	6.3%
Transdermal	187 (12.3)	558 (3.5)	33.5%

References

1. Steele, F. & Diamond, I. Contraceptive switching in Bangladesh. *Stud. Fam. Plann.* **30**, 315–328 (1999).

2. Kungu, W., Agwanda, A. & Khasakhala, A. Prevalence of and factors associated with contraceptive discontinuation in Kenya. *Afr. J. Prim. Health Care Fam. Med.* **14**, 2992 (2022).

3. Grady, W. R., Billy, J. O. & Klepinger, D. H. Contraceptive method switching in the United States. *Perspect. Sex. Reprod. Health* 135–145 (2002).

4. Steinberg, J. R., Marthey, D., Xie, L. & Boudreaux, M. Contraceptive method type and satisfaction, confidence in use, and switching intentions. *Contraception* **104**, 176–182 (2021).

5. Hill, S. *This Is Your Brain on Birth Control: The Surprising Science of Women, Hormones, and the Law of Unintended Consequences*. (Penguin, 2019).

6. Bellizzi, S., Mannava, P., Nagai, M. & Sobel, H. L. Reasons for discontinuation of contraception among women with a current unintended pregnancy in 36 low and middle-income countries. *Contraception* **101**, 26–33 (2020).

7. Daniels, K. & Abma, J. Current contraceptive status among women aged 15–49: United States, 2015–2017. NCHS data brief, no 327. *Natl. Cent. Health Stat.* (2018).

8. Trussell, J. & Vaughan, B. Contraceptive failure, method-related discontinuation and resumption of use: results from the 1995 National Survey of Family Growth. *Fam. Plann. Perspect.* 64–93 (1999).

9. Kavanaugh, M. L. & Jerman, J. Contraceptive method use in the United States: trends and characteristics between 2008, 2012 and 2014. *Contraception* **97**, 14–21 (2018).

10. McDowall, L., Antoniak, M. & Mimno, D. Sensemaking About Contraceptive Methods Across Online Platforms. *ArXiv Prepr. ArXiv230109295* (2023).

11. Ali, M. M. & Cleland, J. Contraceptive switching after method-related discontinuation: levels and differentials. *Stud. Fam. Plann.* **41**, 129–133 (2010).

12. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

13. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).

14. OpenAI. GPT-4 Technical Report. (2023).

15. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are zero-shot clinical information extractors. *ArXiv Prepr. ArXiv220512689* (2022).

16. Goel, A. *et al.* LLMs Accelerate Annotation for Medical Information Extraction. in *Proceedings of the 3rd Machine Learning for Health Symposium* 82–100 (PMLR, 2023).

17. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).

18. Wornow, M. *et al.* The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. *ArXiv Prepr. ArXiv230312961* (2023).

19. Alsentzer, E. *et al.* Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *Npj Digit. Med.* **6**, 1–10 (2023).

20. Wang, M., Sushil, M., Miao, B. Y. & Butte, A. J. Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data. *J. Am. Med. Inform. Assoc.* **30**, 1323–1332 (2023).

21. UCSF Academic Research Systems. UCSF DeID CDW. (2023).

22. Radhakrishnan, L. *et al.* A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open* **6**, ooad045 (2023).

23. Williams, C. Y. K., Miao, B. Y. & Butte, A. J. Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department. 2023.10.19.23297276 Preprint at https://doi.org/10.1101/2023.10.19.23297276 (2023).

24. Liu, P. *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **55**, 195:1-195:35 (2023).

25. Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.* **1**, 309–317 (1957).

26. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr. ArXiv181004805* (2018).

27. Sushil, M., Ludwig, D., Butte, A. J. & Rudrapatna, V. A. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. *ArXiv Prepr. ArXiv221006566* (2022).

28. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* 2623–2631 (2019).

29. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv Prepr. ArXiv220305794* (2022).

30. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* (eds. Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G.) 160–172 (Springer, Berlin, Heidelberg, 2013). doi:10.1007/978-3-642-37456-2_14.

31. Ghassemi, M. *et al.* Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *KDD Proc. Int. Conf. Knowl. Discov. Data Min. Int. Conf. Knowl. Discov. Data Min.*

2014, 75–84 (2014).

32. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
33. Nori, H. *et al.* Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Preprint at https://doi.org/10.48550/arXiv.2311.16452 (2023).
34. Vickery, Z. *et al.* Weight Change at 12 Months in Users of Three Progestin-Only Contraceptive Methods. *Contraception* **88**, 503–508 (2013).
35. Robertson, C. & Braman, A. The New Over-the-Counter Oral Contraceptive Pill — Assessing Financial Barriers to Access. *N. Engl. J. Med.* 389, 1352–1354 (2023).

Chapter 5

Extracting Biologic Treatment Strategies Using Open-Source Language Models

5.1 Introduction

Tumor necrosis factor alpha inhibitors (TNFα-i) are a class of biologic drugs that are used for the treatment of several autoimmune diseases, including inflammatory bowel disease[1,2] (IBD) and rheumatoid arthritis[3,4] (RA). While there are now several biosimilar TNFα-i drugs available for clinical use, there are few biomarkers or clinical guidelines to identify which patient should receive which drug, and treatment failure and switching is common within this class of medications[3,5]. Previous studies have shown that about 14.5% of patients with IBD switch medications at least once and primarily to another TNFα-i[6]. In a cohort of US patients with RA, 39.3% who failed a first-line TNFα-i switched to another TNFα-i[7]. Some patients may develop anti-drug antibodies to TNFα-i, leading to a loss of drug efficacy that may contribute to medication switching[2,8]. Women, older individuals, and patients with high disease activity also tend to have a poor clinical response to TNFα-i and are at higher risk for switching[5,9].

However, these studies often only estimate TNFα-i effectiveness and reasons for switching based on structured medical record data analysis, which often overlook social determinants of health, patient preference, or other reasons for switching[10]. Studies that utilize information from clinical notes require time-consuming manual clinical review, making it difficult to scale and keep these studies up to date. Since mean post-treatment disease activity scores and annual treatment costs tend to be higher in various patients who had a TNFα-i switch[11], there continues to be significant clinical and financial interest to understand the factors driving TNFα-i switching. Here, we aim to develop automated strategies to extract TNFα-i switching information from clinical notes using large language models (LLMs) and to perform a comparative analysis of different open-source LLMs on this task.

5.2 Methods

5.2.1 TNF inhibitor treatment cohort selection

We identified TNFα-i-treated patient cohort using the UCSF Information Commons dataset, which contains longitudinal, deidentified medical record data and clinical notes from patients at a large, academic medical center between 2012 and 2023[12]. We selected all TNFα-i medication orders and administrations, using a string search of all TNFα-i generic or brand names derived from data provided by the Food and Drug Administration[13] (FDA). Medication names were mapped to appropriate generic or biosimilar categories (**Supplemental Table 5.1**), and encounters where a patient switched to a new TNFα-i and had an associated clinical note were identified. Patients without demographic data were also excluded, as were encounters where multiple TNFα-i were ordered on the same date. We also excluded medications without at least 6 months of follow up from downstream comparative analysis, since it could not be determined whether the patient switched TNF inhibitors in these cases. For encounters with multiple notes, only the first note was used for analysis.

Patient demographic information was calculated using the tableone package[14], with continuous distributions reported as means and standard deviations and categorical values represented as proportions.. Statistical testing was performed using chi-square tests for categorical variables and two-sided t-tests for continuous values. A p-value of less than 0.05 was considered significant. All data used in this study was performed using retrospective data, and was determined to be exempt from further review by the UCSF IRB.

5.2.2 Prompt selection for TNFα-i switching reason extraction using GPT-4

The GPT-4 large language model was used to extract information about TNFα-i switching from associated clinical notes in a zero-shot manner. Data were split into 5%/95% validation and test sets, with the validation set used for evaluation of 4 different prompts (**Supplemental Table 5.2**) and final metrics reported on the test set. Each prompt was used to extract the TNFα-i started, the TNFα-i stopped, and one of the following categories for stopping: adverse event, lack of efficacy, insurance/cost, drug resistance, patient preference, other, or unknown ("NA"). Model performance was assessed against weak labels from associated structured medication information, and microF1 scores are reported. Previous experiments have shown that these labels are generally unreliable when assessing null values extracted by GPT-4 (eg. when the medication is not documented in clinical note). As a result, we report microF1 scores separately for all labels and for only non-null values extracted by the language models. The prompt with the highest microF1 scores calculated using non-null values was used to extract TNFα-i switching information from the test set for downstream analysis.

5.2.3 Comparison of open-source large language models on TNFα-i switching information extraction

Several open-source models were also assessed using the manually annotated validation data. These included several independently trained models ("Yi-6B-Chat", "Llama-2-7B-Chat", "Starling-7B-alpha", "Gemma-7B-IT"), as well as updated versions ("Llama-3-8B-Instruct", "Starling-7B-beta") or further finetuned versions of some models ("zephyr-7b-gemma-v0.1", "OpenHermes-2.5-Mistral-7B", "Snorkel-Mistral-PairRM-DPO"). Two models, "JSL-MedMNX-

7B-SFT" and "BioMistral-7B" were specifically trained or finetuned on biomedical data. Additional details on models and parameters usage can be found in the supplemental figures (**Supplemental Table 5.3**). Open source models were compared using average pairwise win rates compared to GPT-4 for each response, following open-source comparative benchmarks[15]. Model "ties" were recorded when both models provide correct or incorrect values, while a model "win" was recorded when one model provides the correct value while the other does not. We report mean win rates of each model against all other models.

5.3 Results

5.3.1 TNFα-i Cohort Identification from UCSF Information Commons

We identified 190,518 relevant TNFα-i medication orders (**Figure 5.1**), including 51,402 that were administered to patients as procedures, from 12,442 unique patients. These orders were mapped to generic names, ignoring dosage information and modality (**Supplemental Table 5.**1). After removing 14 patients without demographic information, 190,500 total medication orders remained. Duplicate TNFα-i orders and orders without associated clinical notes were dropped, leaving 64,983 unique medication orders. When there were different TNFα-i orders at the same encounter, only the first TNFα-i and associated clinical note were considered for downstream analysis. This left a TNFα-i treatment dataset consisting of 58,323 medication orders from 11,572 patients. Of these patients, 2,112 had a documented TNFα-i switch, while 7,075 had no documented switch with a follow-up encounter at least 6 months after the TNFα-i order. Another 2,385 patients also did not have a medication switch but were lost to follow up and were excluded from further analysis.

5.3.2 TNFα-i Treatment Cohort Demographics

The TNFα-i treatment cohort (n=9,187) as a whole had a mean age of 39.9 years (SD 19.0), with a slightly higher proportion of female patients (57.1%, **Table 5.1**). Patients with a documented TNFα-i switch (n=2,112) were more likely to be female (60.5% vs 56.1%, p=0.002) and had a significantly lower mean age (36.3 years, SD 18.4) compared to those who did not (n=7,075, 41.0 years, SD 19.0, p<0.001). Patients with a TNFα-i also tended to be followed longer at UCSF, with a mean follow-up period between their first TNFα-i prescription and final encounter documented at UCSF of 7.1 years (SD 5.6) compared to those without a switch (5.3 years, SD 4.6). There were also significant differences in self-reported race/ethnicity values between the switching and non-switching groups (p<0.001). Overall, the majority of patients were listed as being "White" (60.2%) or "Latinx" (13.9%). There was a higher proportion of patients with a TNFα-i switch who were "Latinx" (16.4%) compared to the non-switching group (13.2%), while there was a lower percentage of patients in the TNFα-i switch group who were "Asian" (7.4% vs 5.8%, respectively).

Proportions of first documented TNFα-i also differed significantly between switching and non-switching groups (p<0.001). The most common first TNFα-i across all patients was Adalimumab (40.9%), which was lower in the TNFα-i switching group (31.6%) compared to the non-switching group (43.7%). Infliximab (26.3%) and etanercept (24.1%) were the next most common, with more patients in the TNFα-i starting on etanercept (31.8%) compared to patients without a TNFα-i switch (21.8%). Within patients who had a TNFα-i switch, the most common first switch was from etanercept to adalimumab (n=546, 6.4%, **Figure 5.2**). Patients who started on certolizumab and infliximab were also most likely to switch to adalimumab (41.8% and 45.5%, respectively). Out of patients who started on adalimumab, the most common switch was to infliximab (n=265, 39.7%).

5.3.3 Reasons for TNFα-i switching using GPT-4 abstracted information

The GPT-4-turbo-128k model was used to test four different prompts for extracting information about TNFα-i switching strategies and reasons for switching (**Supplemental Table 5.2**). Out of the default prompt, prompt that provided specific categories for drug values ("Drugs provided"), a prompt that specified categorical reasons for switching ("Reasons provided"), or both drugs and reason categories provided ("All values provided"), the prompt providing the reason categories had the best overall performance (**Supplemental Table 5.1**). With this prompt, microF1 scores were 0.42 for TNFα-i stopping information extraction and 0.50 for extracting which new TNF was prescribed (n=146). When extracted null values were excluded from analysis, microF1 scores increased to 0.63 (n=71) and 0.89 (n=56), respectively for TNFα-i stopping and new TNFα-i order information. Although all prompts had microF1 scores within 0.05, "Reasons provided" had the highest average score and was used for all downstream tasks.

When applied to the test dataset (n=2958), GPT-4 performance on TNFα-i started and stopped information extraction had microF1 scores of 0.51 and 0.37, respectively. When only considering non-null values, microF1 scores increased to 0.90 (n=1184) and 0.60 (n=1331), respectively. Analysis of all the reasons for TNFα-i switching extracted by GPT-4 for the validation and test sets uncovered that 1759 of the notes appeared to contain no reasons for switching, while the most commonly extracted reason for switching was due to lack of efficacy (n=568, 56.9%, **Supplemental Table 5.4**) and adverse events (n=135, 13.5%). Insurance or cost issues accounted for 10.8% (n=108) of reasons and patient preference for another 8.2% (n=82).

5.3.4 Comparison of TNFα-i information extraction across large language models

The best prompt previously selected ("Reasons provided") was also used to understand how different open source LLMs performed on these treatment information extraction tasks compared to GPT-4. "Starling-7b-beta" had the highest average microF1 score of 0.52 when compared to GPT-4 extracted values while "Llama-2-7B-chat" had the lowest average score of 0.07 (**Supplemental Table 5.5**). Again, only evaluating non-null values increased microF1 scores, which ranged from 0.42 for "Llama-2-7B-chat" to 0.85 for "Starling-7b-beta".

The concordance between models was also explored. Given GPT-4 performance on the previous tasks, outputs from this model were used as a baseline to further evaluate pairwise concordance between other models (**Supplemental Table 5.6**). Llama-3-8B-Instruct and Starling-7B-beta showed the highest mean concordance with GPT-4 extracted information, with a concordance rates of 82.4% (SD: 0.6%) and 77.0% (SD: 6.0%), respectively. Llama-2-7b-Chat showed the lowest concordance, with only 55.3% of values concordant (SD: 7.6%). We also evaluated pairwise win and tie rates of these models compared to GPT-4-turbo-128k extracted values. Mean tie rates ranged from 66.3% (SD: 16.2%, **Figure 5.3**) for llama-2-7b-chat-hf to 80.0% (SD: 12.8%) for JSL-MedMNX-7B-SFT. Llama-3-8B-Instruct had the highest average win rate at 15.5% (SD: 11.4%), followed by zephyr-7b-gemma-v01 at 12.7% (SD: 9.0%).

5.4 Discussion

Here, we provided a set of automated evaluations of GPT-4 and open-source large language models in treatment information extraction from a cohort of TNFα-i-treated patients. We uncovered differences in demographic characteristics for patients who had a TNFα-i switch versus those without, and showed that the most commonly extracted reasons were due to adverse events, followed by lack of efficacy and insurance costs. Adverse events and lack of efficacy are well documented reasons for TNFα-i switching[4,6], and while there have been several studies analyzing the cost-effectiveness of different TNFα-i treatment strategies[11,16], the results here provide evidence for how frequently insurance or cost is a causal reason for TNFα-i switching.

Additionally, we assess the use of GPT-4 and open-source large language models for this use case and showed that prompt engineering with GPT-4 on this task led to comparable performance on many smaller, open-source language models. We further showed that multiple, independent models often can perform this information extraction with highly concordance to GPT-4. As new language models are developed, the need for improved automated evaluation approaches is necessary to understand which models may perform better at different tasks. Finally, we demonstrated that GPT-4 extracted treatment information from notes are often poorly aligned with structured medical record data around medication switching, particularly medication stopping. Future studies are needed to assess whether these discrepancies between structured data and GPT-4 extracted values, as well as the incorporation of non-medical reasons for switching, may affect the development of models for individual treatment estimation or outcome prediction.

There are several limitations to this study. This study did not dive into disease-specific reasons for TNFα-i switching or switches to other medications that may have occurred between TNFα-i switches, although the pipeline described here can be applied to more specific patient cohorts or other classes of medications in future studies. Additionally, our evaluation of open-source language models was only compared to GPT-4, currently the state-of-the-art language model on general benchmarks, and not to expert annotations. Comparisons to expert annotation are likely to clarify the relative capabilities of these open-source models; however there have also been studies showing that human evaluation can also be unreliable and a combination of both evaluations may be beneficial[17,18]. Finally, we did not apply any filters to the data based on time or clinical note type, which may also change the proportions of different reasons extracted.

Despite these limitations, the results presented here contribute insights into both reasons for TNFα-i switching and methods to automate the extraction of such information, using both proprietary and recently developed open-source language models.

5.5 Figures

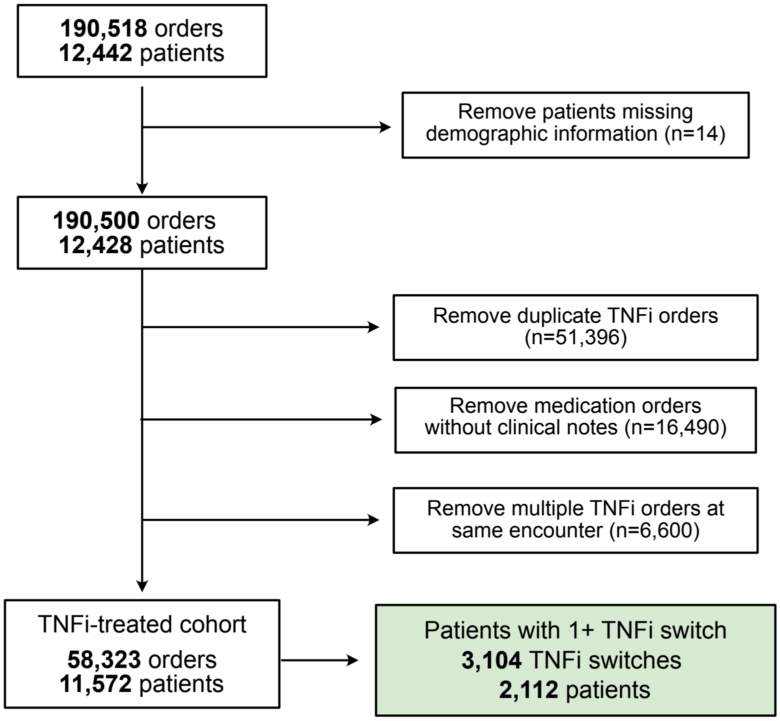

190,518 orders
12,442 patients
Remove patients missing demographic information (n=14)
190,500 orders
12,428 patients
Remove duplicate TNFi orders (n=51,396)
Remove medication orders without clinical notes (n=16,490)
Remove multiple TNFi orders at same encounter (n=6,600)
TNFi-treated cohort
58,323 orders
11,572 patients
Patients with 1+ TNFi switch
3,104 TNFi switches
2,112 patients

Figure 5.1 *Cohort selection.*

TNFα-i-treated patients, and the subset of patients with at least one TNFα-i switch based on medication or procedure orders of relevant drugs, were identified from the UCSF Information Commons dataset.

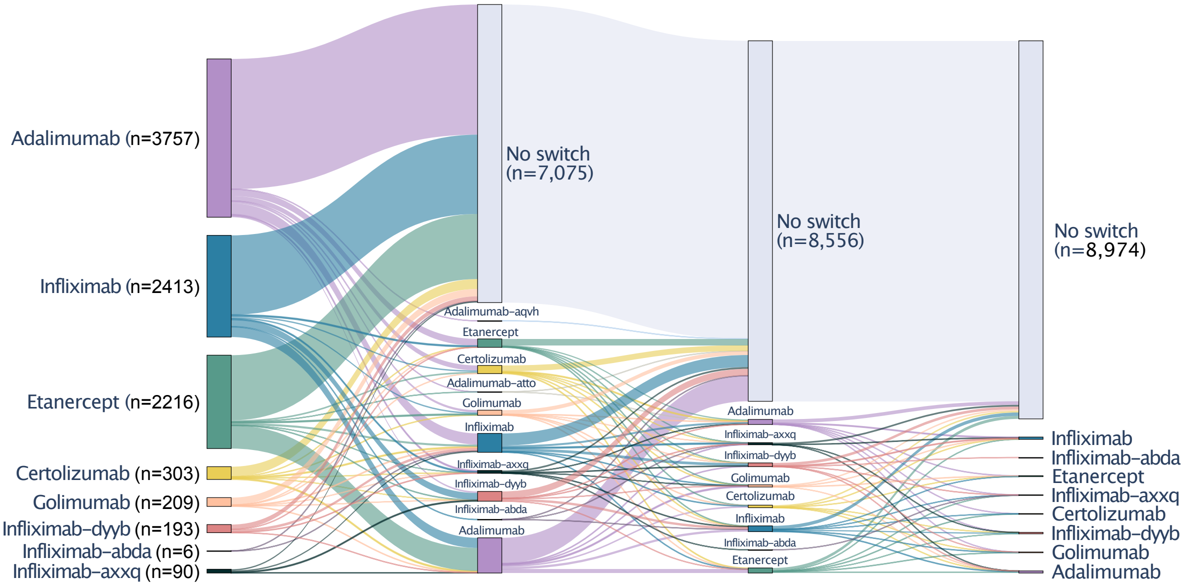


***Figure 5.2** Treatment switching pattern in TNFα-i cohort.*

Sankey diagram showing TNFα-i switching strategies for 9,187 patients from UCSF Information Commons.

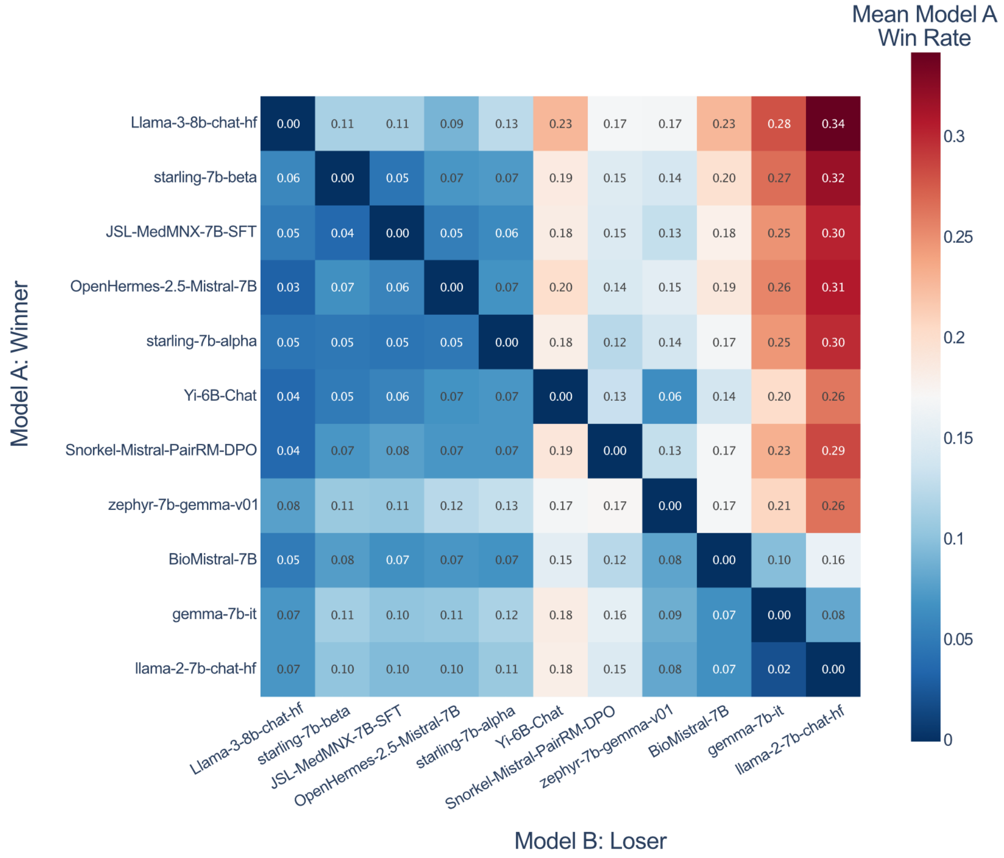


***Figure 5.3** Average win rates across open-source language models compared to GPT-4 extracted TNFα-i switching information.*

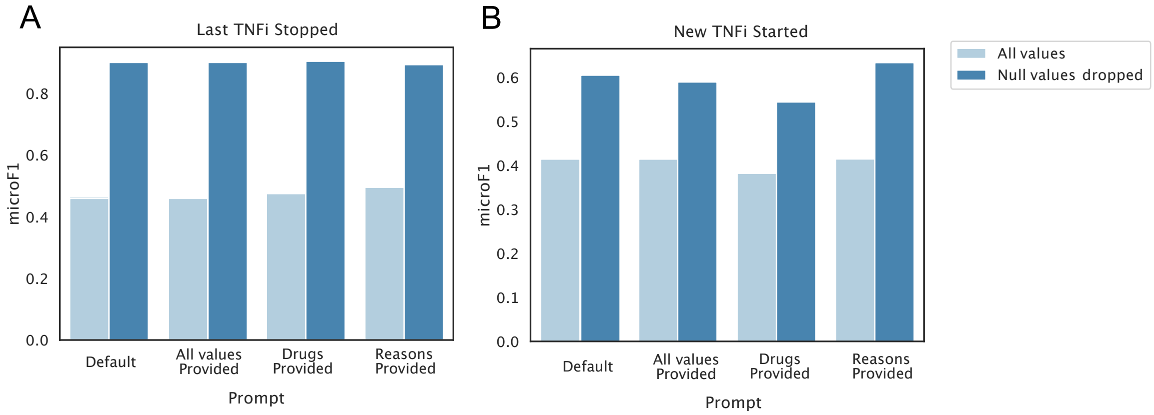
Models were compared pairwise to GPT-4 outputs, with a model “win” occurring if one model matched GPT-4 and the other did not. If both models matched or were discordant, a “tie” was called and not included in this figure.

5.6 Tables

Table 5.1 *Patient demographics.*

	Total (n=9,187)	**No TNFi switch (n=7,705)**	**TNFi switch (n=2,112)**	**Significance**
Mean age, First TNFi (SD)	39.9 (19.0)	41.0 (19.0)	36.3 (18.4)	p<0.001
Mean follow-up, years (SD)	5.7 (4.9)	5.3 (4.6)	7.1 (5.6)	p<0.001
Sex (%)	*Missing = 4*			p=0.002
Female	5244 (57.1)	3969 (56.1)	1275 (60.5)	
Male	3939 (42.9)	3105 (43.9)	834 (39.5)	
Race/Ethnicity (%)	*Missing = 434*			p<0.001
White	5268 (60.2)	4076 (60.9)	1192 (57.7)	
Latinx	1220 (13.9)	881 (13.2)	339 (16.4)	
Other	952 (10.9)	726 (10.9)	226 (10.9)	
Asian	612 (7.0)	492 (7.4)	120 (5.8)	
Black or African American	417 (4.8)	310 (4.6)	107 (5.2)	
Multi-Race/Ethnicity	212 (2.4)	159 (2.4)	53 (2.6)	
Southwest Asian and North African	72 (0.8)	44 (0.7)	28 (1.4)	
First documented TNFi (%)				p<0.001
Adalimumab	3757 (40.9)	3089 (43.7)	668 (31.6)	
Infliximab	2413 (26.3)	1883 (26.6)	530 (25.1)	
Etanercept	2216 (24.1)	1545 (21.8)	671 (31.8)	
Certolizumab	303 (3.3)	236 (3.3)	67 (3.2)	
Infliximab (biosimilar)	289 (3.1)	149 (2.1)	140 (6.6)	
Golimumab	209 (2.3)	173 (2.4)	36 (1.7)	

5.7 Supplemental Figures and Tables



Supplemental Figure 5.1 *Automated evaluation of GPT-4-turbo-128k performance.*

The GPT-4-turbo-128k model was used to extract TNFα-i switching information, including A) which TNFα-i was stopped and B) which was started. microF1 scores evaluated against structured data are shown, with LLM extracted null values included ("All values") and counted as incorrect, or without the null values ("Null values dropped").

Supplemental Table 5.1 *TNF inhibitor drug generic and brand names.*

Brand name	Generic name
cimzia	certolizumab
enbrel	etanercept
humira	adalimumab
remicade	infliximab
simponi	golimumab
eticovo	etanercept-ykro
erelzi	etanercept-szzs
yuflyma	adalimumab-aaty
idacio	adalimumab-aacf
yusimry	adalimumab-aqvh
hulio	adalimumab-fkjp
abrilada	adalimumab-afzb
hadlima	adalimumab-bwwd
hyrimoz	adalimumab-adaz
cyltezo	adalimumab-adbm
amjevita	adalimumab-atto
avsola	infliximab-axxq
ixifi	infliximab-qbtx
renflexis	infliximab-abda
inflectra	infliximab-dyyb

Supplemental Table 5.2 *Prompts tested using automated evaluation.*

Prompt	Prompt Text
Default	Task: Using the clinical note provided, answer the following questions - 1. What new TNF inhibitor (TNFα-i) biologic drug was prescribed or started? If the patient is not starting a new TNF inhibitor drug, write "NA" 2. What was the last TNF inhibitor drug the patient used? If none, write "NA" 3. Why was the last TNF inhibitor drug stopped or planned to be stopped? If no reason was provided or no TNFα-i was stopped, write "NA". Use the following format: {"new_TNFα-i":str, "last_TNFα-i":str, "reason_last_TNFα-i_stopped":str} Answer:
Drugs provided	Task: Cimzia (certolizumab), Enbrel (etanercept), Humira (adalimumab), Remicade (infliximab), Simponi (golimumab), Eticovo (etanercept-ykro), Erelzi (etanercept-szzs), Yuflyma (adalimumab-aaty), Idacio (adalimumab-aacf), Yusimry (adalimumab-aqvh), Hulio (adalimumab-fkjp), Abrilada (adalimumab-afzb), Hadlima (adalimumab-bwwd), Hyrimoz (adalimumab-adaz), Cyltezo (adalimumab-adbm), Amjevita (adalimumab-atto), Avsola (infliximab-axxq), Ixifi (infliximab-qbtx), Renflexis (infliximab-abda), Inflectra (infliximab-dyyb), and Renflixis (infliximab-abda) are tumor necrosis factor inihibitor (TNFα-i) biologic drugs. Using the clinical note provided, extract the following information into this JSON format: {"new_TNFα-i":"What new TNFα-i was prescribed or started? If the patient is not starting a new TNFα-i, write "NA"","last_TNFα-i":"What was the last TNFα-i the patient used? If none, write "NA"","reason_last_TNFα-i_stopped":"Why was the last TNFα-i stopped or planned to be stopped? If no reason was provided or no TNFα-i was stopped, write "NA""} Answer:
Reasons provided	Task: Tumor necrosis factor inhibitors (TNFα-is) describe biologic drugs targeting TNF proteins. Using the clinical note provided, extract the following information into this JSON format: {"new_TNFα-i":"What new TNFα-i was prescribed or started? If the patient is not starting a new TNFα-i, write "NA"","last_TNFα-i":"What was the last TNFα-i the patient used? If none, write "NA"","reason_type_last_TNFα-i_stopped":"Which best describes why the last TNFα-i was stopped or planned to be stopped? "Adverse event", "Drug resistance", "Insurance/Cost","Lack of efficacy","Patient preference","Other", "NA"","full_reason_last_TNFα-i_stopped":"Provide a description for why the last TNFα-i was stopped or planned to be stopped?"} Answer:

Prompt	Prompt Text
All values provided	Task: Cimzia (certolizumab), Enbrel (etanercept), Humira (adalimumab), Remicade (infliximab), Simponi (golimumab), Eticovo (etanercept-ykro), Erelzi (etanercept-szzs), Yuflyma (adalimumab-aaty), Idacio (adalimumab-aacf), Yusimry (adalimumab-aqvh), Hulio (adalimumab-fkjp), Abrilada (adalimumab-afzb), Hadlima (adalimumab-bwwd), Hyrimoz (adalimumab-adaz), Cyltezo (adalimumab-adbm), Amjevita (adalimumab-atto), Avsola (infliximab-axxq), Ixifi (infliximab-qbtx), Renflexis (infliximab-abda), Inflectra (infliximab-dyyb), and Renflixis (infliximab-abda) are tumor necrosis factor inihibitor (TNFα-i) biologic drugs. Using the clinical note provided, extract the following information into this JSON format: {"new_TNFα-i":"What new TNFα-i was prescribed or started? If the patient is not starting a new TNFα-i, write "NA"","last_TNFα-i":"What was the last TNFα-i the patient used? If none, write "NA"","reason_type_last_TNFα-i_stopped":"Which best describes why the last TNFα-i was stopped or planned to be stopped? "Adverse event", "Drug resistance", "Insurance/Cost","Lack of efficacy","Patient preference","Other", "NA"","full_reason_last_TNFα-i_stopped":"Provide a description for why the last TNFα-i was stopped or planned to be stopped?"} Answer:

Supplemental Table 5.3 *Open source model information.*

Model name	**Base Model**	**Release date**	**Reference or model repository**
Llama-2-7b-Chat	Llama-2-7B	Jul 18, 2023	Touvron et al, 2023[19]
OpenHermes-2.5-Mistral-7B	Mistral-7B	Oct 29, 2023	https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B
Yi-6B-Chat	Yi-6B	Nov 22, 2023	01.AI 2024[20]
Starling-7b-alpha	Mistral-7B	Nov 25, 2023	Zhu et al, 2023[21]
Snorkel-Mistral-PairRM-DPO	Mistral-7B	Jan 19, 2024	https://huggingface.co/snorkelai/Snorkel-Mistral-PairRM-DPO
BioMistral-7B	Mistral-7B	Feb 14, 2024	Labrak et al, 2024[22]
Gemma-7B-IT	Gemma-7B	Feb 21, 2024	https://huggingface.co/google/gemma-7b-it
Zephyr-7b-gemma-v01	Gemma-7B	Mar 1, 2024	https://huggingface.co/HuggingFaceH4/zephyr-7b-gemma-v0.1
Starling-7b-beta	Mistral-7B	Mar 19, 2024	https://huggingface.co/Nexusflow/Starling-LM-7B-beta
JSL-MedMNX-7B-SFT	Starling-7B (?)	April 15, 2024	https://huggingface.co/johnsnowlabs/JSL-MedMNX-7B-SFT
Llama-3-8b-Instruct	Llama-3-8B	April 18, 2024	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Supplemental Table 5.4 *Reasons for treatment switching.*

Reason type	Count	Proportion
NA/Unknown	2106	-
Lack of efficacy	568	56.9%
Adverse event	135	13.5%
Insurance/Cost	108	10.8%
Patient preference	82	8.2%
Other	54	5.4%
Drug resistance	51	5.1%

Supplemental Table 5.5 *Open-source language model performance on validation set.*

Model	Prompt	Extraction Task	All values (microF1)	Null values dropped (microF1)	Non-null values (n)
gpt-4-turbo-128k	default-task	TNFα-i Stopped	0.41	0.61	76
gpt-4-turbo-128k	default-task	TNFα-i Started	0.46	0.88	52
gpt-4-turbo-128k	all-values-provided	TNFα-i Stopped	0.41	0.59	78
gpt-4-turbo-128k	all-values-provided	TNFα-i Started	0.46	0.90	50
gpt-4-turbo-128k	drugs-provided	TNFα-i Stopped	0.38	0.54	79
gpt-4-turbo-128k	drugs-provided	TNFα-i Started	0.47	0.90	52
gpt-4-turbo-128k	reasons-provided	TNFα-i Stopped	0.41	0.63	71
gpt-4-turbo-128k	reasons-provided	TNFα-i Started	0.50	0.89	56
starling-7b-beta	reasons-provided	TNFα-i Stopped	0.43	0.83	52
starling-7b-beta	reasons-provided	TNFα-i Started	0.60	0.88	76
llama-3-8b-chat-hf	reasons-provided	TNFα-i Stopped	0.44	0.79	56
llama-3-8b-chat-hf	reasons-provided	TNFα-i Started	0.57	0.83	76
JSL-MedMNX-7B-SFT	reasons-provided	TNFα-i Stopped	0.41	0.82	49
JSL-MedMNX-7B-SFT	reasons-provided	TNFα-i Started	0.59	0.83	81
OpenHermes-2.5-Mistral-7B	reasons-provided	TNFα-i Stopped	0.38	0.88	42
OpenHermes-2.5-Mistral-7B	reasons-provided	TNFα-i Started	0.61	0.82	85
starling-7b-alpha	reasons-provided	TNFα-i Stopped	0.37	0.87	39
starling-7b-alpha	reasons-provided	TNFα-i Started	0.60	0.82	84
Yi-6B-Chat	reasons-provided	TNFα-i Stopped	0.39	0.63	67
Yi-6B-Chat	reasons-provided	TNFα-i Started	0.57	0.83	77
Snorkel-Mistral-PairRM-DPO	reasons-provided	TNFα-i Stopped	0.40	0.76	51
Snorkel-Mistral-PairRM-DPO	reasons-provided	TNFα-i Started	0.54	0.77	79

Model	**Prompt**	**Extraction Task**	**All values (microF1)**	**Null values dropped (microF1)**	**Non-null values (n)**
zephyr-7b-gemma-v01	reasons-provided	TNFα-i Stopped	0.34	0.56	64
zephyr-7b-gemma-v01	reasons-provided	TNFα-i Started	0.46	0.94	48
BioMistral-7B	reasons-provided	TNFα-i Stopped	0.22	0.86	21
BioMistral-7B	reasons-provided	TNFα-i Started	0.38	0.80	45
gemma-7b-it	reasons-provided	TNFα-i Stopped	0.03	1.00	2
gemma-7b-it	reasons-provided	TNFα-i Started	0.05	0.89	5
llama-2-7b-chat-hf	reasons-provided	TNFα-i Stopped	0.04	0.22	14
llama-2-7b-chat-hf	reasons-provided	TNFα-i Started	0.10	0.62	13

Supplemental Table 5.6 *Concordance of open-source language models with GPT-4 values.*

	Previous TNFα-i	**TNFα-i Started**	**Reason Stopped**	**Mean**	**SD**
gpt-4-turbo-128k	100.0%	100.0%	100.0%	100.0%	0.0%
llama-3-8b-chat-hf	82.9%	81.5%	82.9%	82.4%	0.6%
starling-7b-beta	80.8%	81.5%	68.5%	76.9%	6.0%
JSL-MedMNX-7B-SFT	78.1%	80.1%	69.9%	76.0%	4.4%
OpenHermes-2.5-Mistral-7B	72.6%	75.3%	81.5%	76.5%	3.7%
starling-7b-alpha	72.6%	76.7%	74.0%	74.4%	1.7%
Yi-6B-Chat	68.5%	75.3%	45.9%	63.2%	12.6%
Snorkel-Mistral-PairRM-DPO	73.3%	69.9%	64.4%	69.2%	3.7%
zephyr-7b-gemma-v01	71.9%	83.6%	65.1%	73.5%	7.6%
BioMistral-7B	59.6%	73.3%	61.0%	64.6%	6.2%
gemma-7b-it	52.7%	65.1%	66.4%	61.4%	6.2%
llama-2-7b-chat-hf	44.5%	61.0%	60.3%	55.3%	7.6%

References

1. Rudrapatna, V. A. & Velayos, F. Biosimilars for the Treatment of Inflammatory Bowel Disease. *Pract. Gastroenterol.* **43**, 84–91 (2019).

2. Atiqi, S., Hooijberg, F., Loeff, F. C., Rispens, T. & Wolbink, G. J. Immunogenicity of TNF-Inhibitors. *Front. Immunol.* **11**, 312 (2020).

3. Fraenkel, L. *et al.* 2021 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. *Arthritis Rheumatol.* **73**, 1108–1123 (2021).

4. Tesser, J. *et al.* Improvement in disease activity among patients with rheumatoid arthritis who switched from intravenous infliximab to intravenous golimumab in the ACR RISE registry. *Clin. Rheumatol.* **41**, 2319–2327 (2022).

5. Law-Wan, J. *et al.* Predictors of response to TNF inhibitors in rheumatoid arthritis: an individual patient data pooled analysis of randomised controlled trials. *RMD Open* **7**, e001882 (2021).

6. Meijboom, R. W. *et al.* Switching TNFα inhibitors: Patterns and determinants. *Pharmacol. Res. Perspect.* **9**, e00843 (2021).

7. Caporali, R., Conti, F. & Iannone, F. Management of patients with inflammatory rheumatic diseases after treatment failure with a first tumour necrosis factor inhibitor: A narrative review. *Mod. Rheumatol.* **34**, 11–26 (2024).

8. Bellur, S. *et al.* Antidrug Antibodies to Tumor Necrosis Factor α Inhibitors in Patients With Noninfectious Uveitis. *JAMA Ophthalmol.* **141**, 150–156 (2023).

9. Yoosuf, N. *et al.* Early prediction of clinical response to anti-TNF treatment using multi-omics and machine learning in rheumatoid arthritis. *Rheumatology* **61**, 1680–1689 (2022).

10. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *Npj Digit. Med.* **7**, 1–14 (2024).

11. Song, Y. *et al.* Economic Burden of Switching to Different Biologic Therapies Among TNFi-Experienced Patients with Psoriatic Arthritis. *Rheumatol. Ther.* **6**, 285–297 (2019).

12. Radhakrishnan, L. *et al.* A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open.* **6**, ooad045 (2023).

13. Biosimilar Product Information. *FDA* (2023). Last accessed May 2, 2024.

14. Pollard, T. J., Johnson, A. E. W., Raffa, J. D. & Mark, R. G. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open* **1**, 26–31 (2018).

15. Zheng, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Preprint at https://doi.org/10.48550/arXiv.2306.05685 (2023).

16. Cannon, G. W. *et al.* Clinical Outcomes and Biologic Costs of Switching Between TNFi in US Veterans with Rheumatoid Arthritis. *Adv. Ther.* **33**, 1347–1359 (2016).

17. Hosking, T., Blunsom, P. & Bartolo, M. Human Feedback is not Gold Standard. Preprint at https://doi.org/10.48550/arXiv.2309.16349 (2024).

18. Sylolypavan, A., Sleeman, D., Wu, H. & Sim, M. The impact of inconsistent human annotations on AI driven clinical decision making. *Npj Digit. Med.* **6**, 1–13 (2023).

19. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).

20. AI.01 *et al.* Yi: Open Foundation Models by 01.AI. Preprint at https://doi.org/10.48550/arXiv.2403.04652 (2024).

21. Zhu, B., Frick, E., Wu, T., Zhu, H. & Jiao, J. Starling-7B: Increasing LLM Helpfulness & Harmlessness with RLAIF. https://starling.cs.berkeley.edu (2023).

22. Labrak, Y. *et al.* BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. Preprint at https://doi.org/10.48550/arXiv.2402.10373 (2024).

Chapter 6

Generation of Guideline-Based Clinical Decision Trees

6.1 Abstract

Molecular biomarkers play a pivotal role in the diagnosis and treatment of oncologic diseases but staying updated with the latest guidelines and research can be challenging for healthcare professionals and patients. Large Language Models (LLMs), such as MedPalm-2 and GPT-4, have emerged as potential tools to streamline biomedical information extraction, but their ability to summarize molecular biomarkers for oncologic disease subtyping remains unclear. Auto-generation of clinical nomograms from text guidelines could illustrate a new type of utility for LLMs.

In this cross-sectional study, two LLMs, GPT-4 and Claude-2, were assessed for their ability to generate decision trees for molecular subtyping of oncologic diseases with and without expert-curated guidelines. Clinical evaluators assessed the accuracy of biomarker and cancer subtype generation, as well as validity of molecular subtyping decision trees across five cancer types: colorectal cancer, invasive ductal carcinoma, acute myeloid leukemia, diffuse large B-cell lymphoma, and diffuse glioma. Both GPT-4 and Claude-2 "off the shelf" successfully produced clinical decision trees that contained valid instances of biomarkers and disease subtypes.

Overall, GPT-4 and Claude-2 showed limited improvement in the accuracy of decision tree generation when guideline text was added. A Streamlit dashboard [https://clinicaltrees.org/] was developed for interactive exploration of subtyping trees generated for other oncologic diseases. This study demonstrates the potential of LLMs like GPT-4 and Claude-2 in aiding the summarization of molecular diagnostic guidelines in oncology. While effective in certain aspects,

their performance highlights the need for careful interpretation, especially in zero-shot settings. Future research should focus on enhancing these models for more nuanced and probabilistic interpretations in clinical decision-making. The developed tools and methodologies present a promising avenue for expanding LLM applications in various medical specialties.

6.2 Introduction

Molecular biomarkers are becoming increasingly crucial in supporting the diagnosis and treatment of oncologic diseases but keeping up with the latest guidelines and relevant research can be time-consuming for physicians, researchers, and patients. The recent emergence of several new large language models (LLMs) present a unique opportunity to help streamline text-heavy healthcare workflows, including medical information summarization and education. Previous studies have demonstrated that new LLMs are capable of extracting complex clinical information from oncology progress notes[1], suggesting differential diagnoses[2], or even generating decision trees from clinical trial criteria[3] or for clinical decision support[4]. The generation of decision trees can provide clear visual guidelines for clinical support, which can significantly impact downstream clinical care. In this study, we aimed to assess the capabilities of two recently developed LLMs in generating diagnostic decision trees for the molecular subtyping of cancers, using published clinical guidelines.

6.3 Methods

Diagnostic trees describing cancer subtypes based on molecular biomarker status were generated for five cancers using GPT-4 (OpenAI) and Claude-2 (Anthropic), two LLMs with public Application Programming Interfaces (APIs). These cancers were selected based on the prevalence of known molecular biomarkers, and included two common solid organ cancers (colorectal cancer [CRC] and invasive ductal carcinoma [IDC]), a common hematologic cancer (acute myeloid leukemia, AML), a rare hematologic cancer (diffuse large B-cell lymphoma [DLBCL]), and a rare solid cancer (diffuse glioma).

Trees were generated using a specific prompt that contained either only formatting guidelines (**Figure 6.1**) or also included information provided from recent classification guidelines for each of the five cancers[5–9] (**Supplemental Table 6.1**). Clinical trees were generated to contain molecular biomarker status as nodes, terminating at nodes that were molecular subtypes. Model temperature was set to 0, and a new API call was made for each of the different prompts used. Additional details on models and parameters used are provided in **Supplemental Table 6.2**. Results were processed into Pydot graph objects[10] and visualized using an interactive dashboard developed using Streamlit[11].

Each branch of LLM-generated decision trees were evaluated against subtyping decision trees generated by clinical reviewers based on clinical guidelines. Evaluators were blinded to which language model generated which tree, and each tree was evaluated by two reviewers, with discrepancies resolved by discussion. We report mean accuracies of subtyping trees, as well as proportions of subtypes and biomarkers correctly extracted by the two LLMs for each cancer. Hallucinations, identified as values not mentioned in recent guidelines for use in molecular cancer subtype diagnosis, were also quantified by clinical evaluators. Accuracy of LLM trees with and without guidelines were compared with two-sided T-tests using Scipy[12]. A p-value less than 0.05 was considered statistically significant.

6.4 Results

Both Claude-2 and GPT-4 were able to create properly formatted decision trees with or without being provided actual clinical guideline text. Including guideline text improved the proportion of cancer subtypes and biomarkers that each model was able to extract. Mean accuracy of cancer subtype extraction increased when guidelines were provided, with the Claude-2 model increasing from 45% (SD: 44.7%, n=5) to 81.9% (SD: 20.8%, p=0.13) and GPT-4 from 36.1% (SD: 33.3%) to 82.0% (SD: 24.2%, p=0.035). Without guidelines, both GPT-4 and Claude-2 were best at generating accurate cancer subtypes in decision trees for IDC (80% and 100%, respectively) and neither were able to produce subtypes of CRC. By providing guideline text, both GPT-4 and Claude-2 were able to extract and visualize all expected subtypes for IDC and CRC (**Supplemental Figure 6.1**).

Regarding hallucinations, GPT-4 and Claude-2 produced the greatest proportion of hallucinated subtypes, which were subtypes not present in clinical trees generated by clinical annotators, for CRC and AML when not provided guideline text. Subtypes that were not mentioned in recent guidelines, such as "NPM1 Wildtype, FLT3-ITD Wildtype and CEBPA Mutated AML," were considered hallucinations. On average, 40% (SD: 54.8%) of subtypes extracted by Claude-2 without guidelines were deemed to be hallucinations, which decreased to 21.0% (SD: 23.7%, p=0.50) when provided guideline text . GPT-4 referenced hallucinated cancer subtypes 37.1% (SD: 54.8%) of the time when not provided guideline text, which dropped to 2.9% (SD: 6.3%, p=0.17) when provided with guideline text (**Supplemental Figure 6.1**).

For accurate biomarker extraction, Claude-2 extracted 55.3% of expected biomarkers on average (SD: 24.6%) without guideline text and 86.2% with (SD: 16.4% , p=0.07), while GPT-4 extracted 50.3% (SD: 27.3%) of biomarkers without guideline text and 83.3% with (SD: 23.5%, p=0.048). Without guideline text, both GPT-4 and Claude-2 both showed 75% accuracy for biomarker extraction for IDC and were least accurate in extracting biomarkers for AML (4.2% and 12.5%, respectively). With guideline text, both GPT-4 and Claude-2 were able to extract all expected subtypes for IDC and diffuse gliomas (**Supplemental Figure 6.2**).

On average, without guideline text, Claude-2 and GPT-4 produced biomarkers that were considered hallucinations (for example, "RBM15::MKL1" and "TP53") in 16.3% (SD: 17.1%) and 16.0% (SD: 35.8%) of generated values, respectively. With provided guideline text, hallucinations decreased to 12.5% (SD: 13.6%) for Claude-2 and 13.0% (SD: 15.9%) for GPT-4. The largest proportion of hallucinated biomarkers was produced for AML, with 40% hallucinations for Claude-2 and 80% for GPT-4, although providing guidelines reduced model hallucination down to 8.7% for Claude-2 and 7.7% for GPT-4 (**Supplemental Figure 6.2**).

Assessment of average overall accuracy of decision trees showed that, without guidelines, GPT-4 produced valid branches 46.7% (SD: 46.2%) of the time, while decision tree branches were 39.3% (SD: 40.1%) valid for Claude-2. Substantial increases in decision tree accuracy were seen for AML, going from 0% to 92.3% for GPT-4 and 0% to 61.7% for Claude-2. However, adding in guideline text did not significantly increase overall accuracy of decision tree generation for either GPT-4, which increased to 72.5% (SD: 41.1%, p=0.38) or Claude-2 (54.2%, SD: 30.5%, p=0.52).

A streamlit dashboard [https://clinicaltrees.org/] was developed to provide a user interface for exploration of GPT-4 and Claude-2 model performance on subtyping tree extraction for user-specified cancer types and guidelines (**Figure 6.3**).

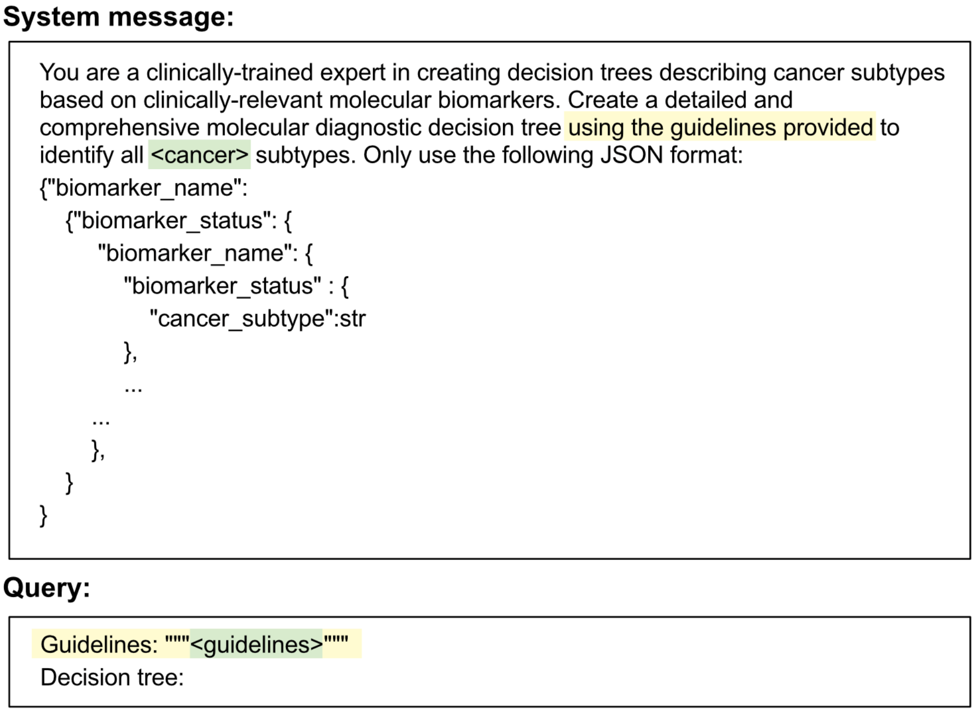
6.5 Discussion

Here, we demonstrate the capability for language models to generate accurate and comprehensive decision trees from clinical guideline text for molecular diagnosis across multiple cancer types. Additionally, we showed that adding clinical guideline text into prompts improves extraction of molecular biomarkers and oncology disease subtypes but did not significantly improve clinical decision tree generation.

While this brief report identifies opportunities for LLMs in supporting biomedical information review and visualization in oncology, the results are focused on molecular diagnosis, which is only a part of clinical decision making. Furthermore, not all molecular features are binary in nature, and future iterations of these decision trees may be assessed for their ability to include probabilities at each branch along the decision tree. Finally, another limitation to this study is the use of API-based models, which are not as interpretable and are more costly to run compared to open-source alternatives. We also did not perform any prompt engineering, and further exploration of strategies like chain-of-thought may help improve decision tree generation, which involves significant reasoning capabilities.

Despite these limitations, our initial evaluation of GPT4 for oncology molecular information extraction shows significant potential for further development. Additionally, we provide open access to the tools assessed and developed here, and for future studies to use similar approaches to evaluate summarization of guidelines for treatment or other aspects of clinical workflows across different medical specialties. Future work might even include being able to summarize many raw clinical studies and results from clinical trials into more accessible guideline texts and visualizations.

6.6 Figures



System message:

You are a clinically-trained expert in creating decision trees describing cancer subtypes based on clinically-relevant molecular biomarkers. Create a detailed and comprehensive molecular diagnostic decision tree using the guidelines provided to identify all <cancer> subtypes. Only use the following JSON format:

```
{"biomarker_name":
   {"biomarker_status": {
        "biomarker_name": {
            "biomarker_status" : {
                "cancer_subtype":str
            },
            ...
        ...
        },
   }
}
```

Query:

Guidelines: """<guidelines>"""
Decision tree:

Figure 6.1 *Prompts to generate clinical decision trees.*

Prompt used to generate clinical cancer subtyping trees. Values highlighted in green are replaced with cancer specific information for each of the five cancers evaluated, and values highlighted in yellow are only included if guidelines are present.

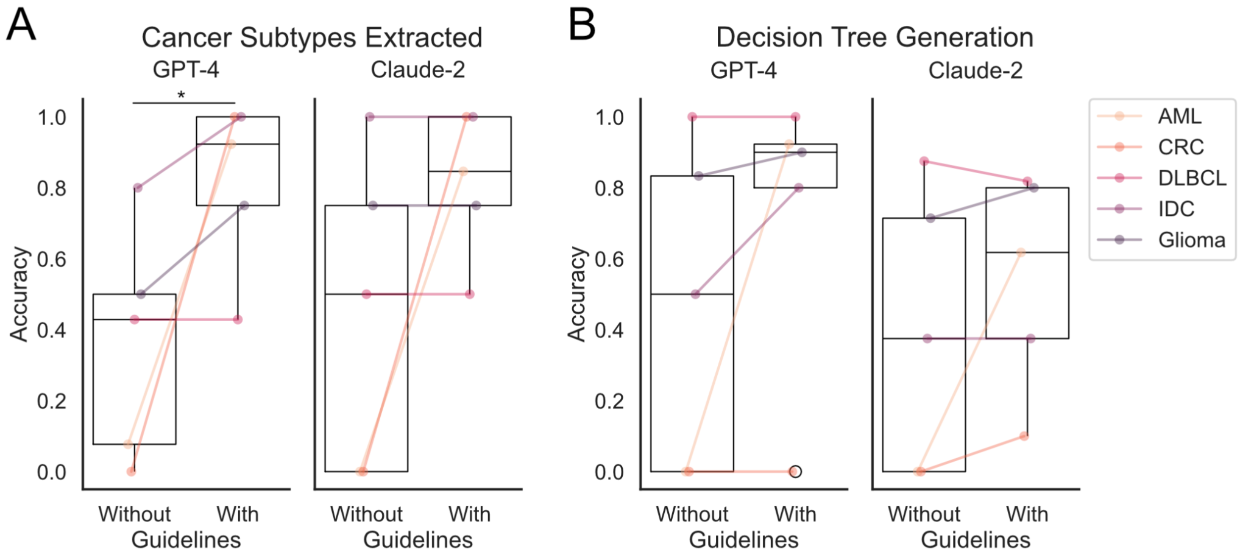
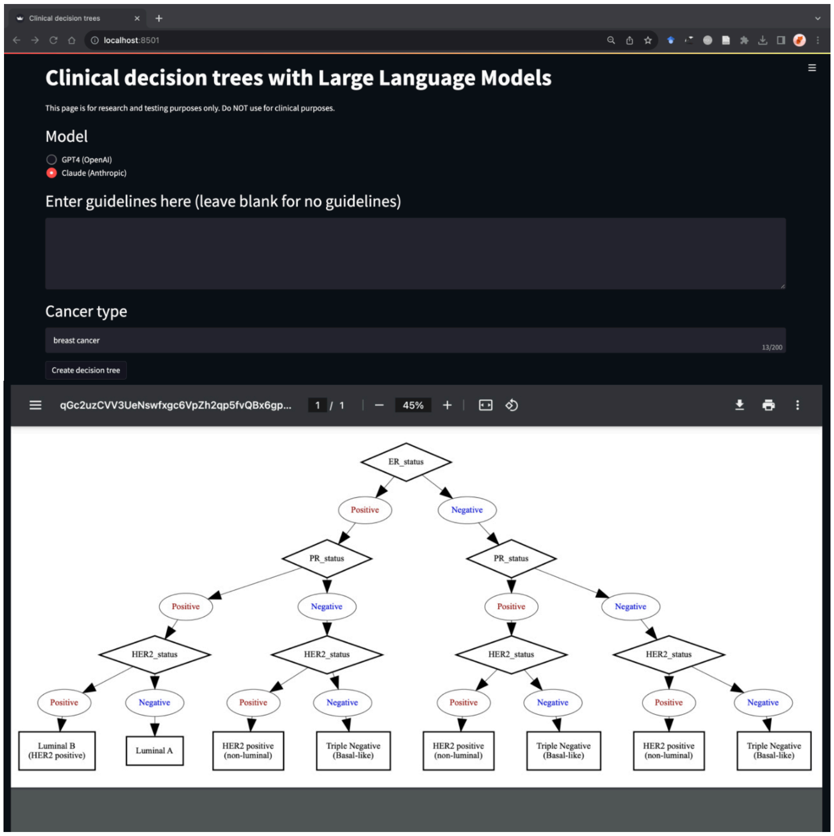


Figure 6.2 *Accuracy of clinical decision tree generation using LLMs.*

Clinical evaluators assessed the A) accuracy of cancer subtype extracted by each LLM with and without guidelines. B) Clinical evaluators also assessed the overall accuracy of clinical decision trees generated. A tree was only considered correct if all biomarkers and subtypes were clinically appropriate, and the biomarkers accurately described the associated cancer subtype.



Clinical decision trees with Large Language Models

This page is for research and testing purposes only. Do NOT use for clinical purposes.

Model

○ GPT4 (OpenAI)
● Claude (Anthropic)

Enter guidelines here (leave blank for no guidelines)

Cancer type

breast cancer 13/200

Create decision tree

Figure 6.3 *Clinical decision tree dashboard.*

A streamlit dashboard [https://clinicaltrees.org/] was created to enable exploration of subtyping decision trees for other cancers and guidelines.

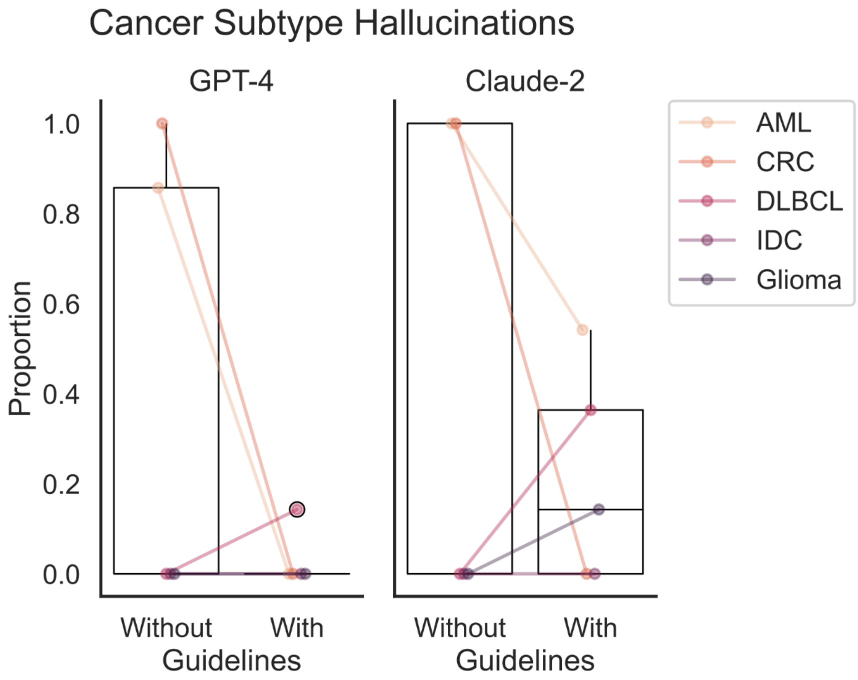
6.7 Supplemental Figures and Tables

Supplemental Table 6.1 *Guideline references and sections used.*

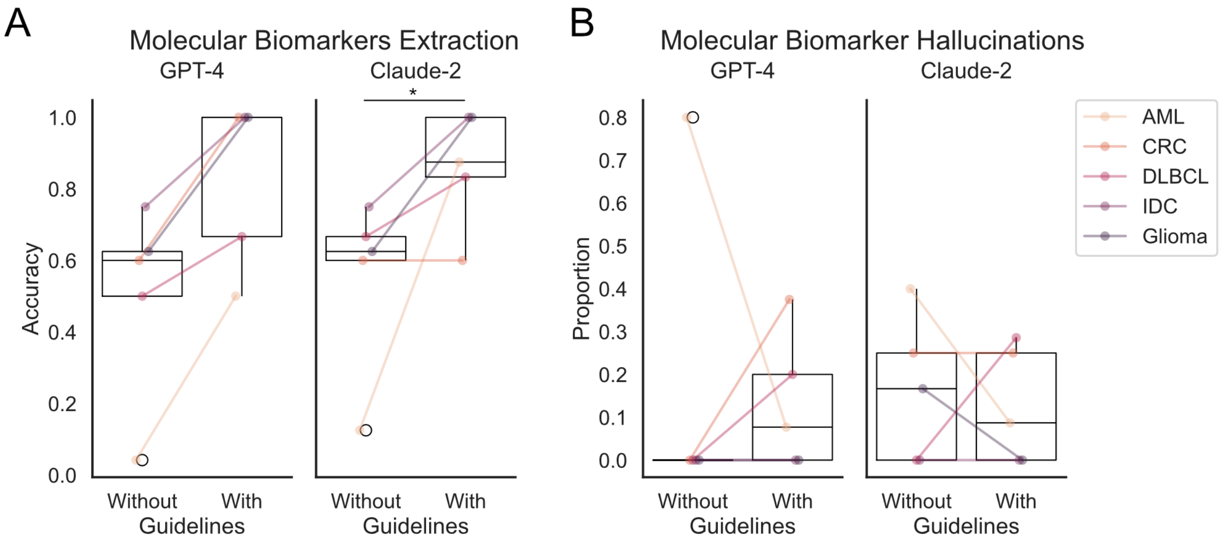
Cancer type and reference used	Sections used
Acute Myeloid Leukemia[6]	"Acute myeloid leukaemia" section 1 (Enhanced grouping framework permitting scalable genetic classification and deemphasizing blast enumeration where relevant) and section 2 (AML with defining genetic abnormalities)
Diffuse large B-cell lymphoma[5]	"Overview," "Diagnosis," and "Workup"
Diffuse gliomas[8]	"Integrated histomolecular classification"
Colorectal cancer[7]	"CMS1," "CMS2," "CMS3," and "CMS4"
Invasive ductal carcinoma[9]	"Molecular classification" section 1 (Intrinsic Subtypes) and section 2 (Integrative Clusters)

Supplemental Table 6.2 *Overview of language models used.*

	GPT4	**Claude**
Company	OpenAI	Anthropic
Model name	GPT4	Claude 2
Model version	0613	2023-06-01
Model context length	8,192	100,000
Temperature	0	0
Top p	1	1
Maximum output tokens	4500	4500
Training data cutoff date	September 2021	December 2022
Cost	$0.03/1k input tokens $0.06/1k output tokens	$11.02/1M input tokens $32.68/1M output tokens
Date accessed	October 3, 2023	October 3, 2023



***Supplemental Figure 6.1** Hallucinations of cancer subtype extraction*



Supplemental Figure 6.2 *Accuracy and hallucinations of cancer subtype extraction*

References

1. Sushil, M. e*t al*. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 1(4):AIdbp2300110 (2024).

2. Benary, M. *et al*. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*. 6(11):e2343689- (2023).

3. Wong, C. *et al*. Scaling clinical trial matching using large language models: A case study in oncology. *PMLR*. 219:846-862 (2023).

4. Zhu, W., Li, W., Tian, X., et al. Text2MDT: Extracting Medical Decision Trees from Medical Texts. *arXiv*. https://doi.org/10.48550/arXiv.2401.02034 (2024). (preprint).

5. Li, S., Young, K. H. & Medeiros, L. J. Diffuse large B-cell lymphoma. *Pathology (Phila)*. 50(1):74-87 (2018).

6. Khoury, J. D. *et al*. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours. *Leukemia*. 36(7):1703-1719 (2022).

7. Thanki, K. *et al*. Consensus Molecular Subtypes of Colorectal Cancer and their Clinical Implications. *Int Biol Biomed J*. 3(3):105-111 (2017).

8. Weller, M. *et al*. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nat Rev Clin Oncol*. 18(3):170-186 (2021).

9. Tsang, J. Y. S. & Tse, G.M. Molecular Classification of Breast Cancer. *Adv Anat Pathol*. 27(1):27-35 (2020).

10. pydot. Accessed February 12, 2024. https://pypi.org/project/pydot/

11. Streamlit Docs. Accessed February 12, 2024. https://docs.streamlit.io/

12. Virtanen, P. *et al*. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17(3):261-272 (2020).

Chapter 7

Checklist for Generative Modeling for Clinical Research

Recent advances in generative models, including large language models (LLMs), vision language models (VLMs), and diffusion models, have accelerated the field of natural language and image processing in medicine and marked a significant paradigm shift in how biomedical models can be developed and deployed[1,2]. While these models are highly adaptable to new tasks, scaling and evaluating their usage presents new challenges not addressed in previous frameworks. In particular, the ability of these models to produce useful outputs with little to no specialized training data ("zero-" or "few-shot" approaches), as well as the open-ended nature of their outputs, necessitate the development of updated guidelines in using and evaluating these models.

In response to gaps in standards and best practices for the development of clinical AI tools identified by US Executive Order 14110[3] and several emerging national networks for clinical AI evaluation[4], we begin to formalize some of these guidelines by building on the "Minimum information about clinical artificial intelligence modeling" (MI-CLAIM) checklist[5].

The MI-CLAIM checklist, originally developed in 2020 and already cited 300 times, provided a set of six steps with guidelines on the minimum information necessary to encourage transparent, reproducible research for artificial intelligence (AI) in medicine. Here, we propose modifications to the original checklist that highlight differences in training, evaluation, interpretability, and reproducibility of generative models compared to traditional AI models for clinical research. This updated checklist also seeks to clarify cohort selection reporting and adds additional items on alignment with ethical standards.

7.1 Study design

We describe best-practice approaches for generative modeling study design, particularly for new tasks enabled by these technologies and how these may affect study design choices. Additionally, we update the MI-CLAIM to clarify checklist items to improve reproducibility of cohort selection for all clinical AI research. Finally, we add checklist items to encourage researchers to assess bias, privacy, and harm of generative AI studies (**Table 7.1**).

7.1.1 Study design for generative modeling

Generative modeling has opened up new types of tasks that have previously been limited by the capabilities of older models, and require careful consideration of appropriate datasets, labels, evaluation, and interpretation of results. In tasks where the outcomes are well-defined and fall into either discrete categories (categorical labels) or a spectrum of values (continuous labels), these labels should be robust, clinically validated, and reflective of the outcome of interest. How labels were derived should also be clearly documented, including the source of the labels and protocols to retrieve these labels. If labels are provided by human annotators, at least two annotators should be involved and details on annotation guidelines and inter-annotator agreement provided[6]. If outputs are unstructured, such as summaries of clinical notes, and do not readily map to simple labels, more robust evaluation frameworks are necessary, which may again involve both automated and human evaluation. We discuss these evaluation strategies in detail in part 4.

For all generative model studies, researchers should also be careful of training data memorization (“data leakage” or “contamination”)[7,8]. Almost all publicly available datasets are included in generative model training data and should not be used as test datasets unless it can be demonstrated that the model has not been trained on the specific task or the data was published after the model was trained[9]. One option to test for memorization is to see whether the generative model can regenerate large portions of the dataset[10]. Importantly, however, memorization of data is still possible even if the foundation model cannot regenerate the dataset in this way and should be listed as a limitation if public datasets are used[11].

Another key feature of generative models is their ability to produce a variety of valid outputs, making consistency and reproducibility a unique challenge in study design. For classification problems where the goal is to consistently output a discrete value, models can be set to a temperature of 0, greedy sampling to select the highest probability token each time, or a seed set to control reproducibility[6]. This is particularly relevant in clinical decision support systems where consistency of recommendations is a key concern. For open-ended generation, where stochasticity is expected or of interest to the research, sampling confidence intervals or representative outputs should be provided where feasible[10,11].

7.1.2 Best practices for cohort selection

Ideally, code to select patient cohorts and raw individual-level data should be made available (which is increasingly compliant with mandates from funding agencies, including the National Institutes of Health), but in cases where either is not possible, full details on both the patient cohort selection should be provided. Ambiguous language, such as "patients diagnosed with diabetes were included," should be avoided in favor of more reproducible terms, such as "patients who had at least 2 of the following ICD-10 codes: E11.*, E13.*... were included". If datasets are de-identified or are otherwise not representative of the clinical settings presented by the research question, these limitations should be described and discussed in detail. This information may include how date were shifted to preserve privacy, whether age is masked, specific methods used for redaction of text, which Electronic Medical Record (EMR) vendor the data was derived from, if the data was obtained from specific department(s), or other deviations from real-world settings.

We additionally provide checklist items for cohort selection based on unstructured or multimodal data. If methods to select patients are based on the presence of certain values mentioned in clinical text, the list of keyword terms, regular expressions, or other selection criteria should be made available. If qualitative factors, such as manual chart review, are used to identify patient cohorts, these should be detailed and the qualifications (eg. years of practice, specialty, etc) of the reviewer should be reported. Pre- and post-processing steps, such as extracting specific sections, converting text to lowercase, lemmatization or stemming, and/or mapping to standard vocabularies, should also be reported in full. Sensitivity analyses should be performed where appropriate to justify any patient selection criteria deviating from established guidelines. Specifications for handling missing data should also be provided, if applicable.

7.1.3 Bias, privacy, and harm assessments

Identifying potential harms of modeling approaches is becoming increasingly important for generative models, which can produce complex, unstructured outputs that may be difficult to identify as inaccurate or biased[12,13]. The updated MI-CLAIM checklist introduces new items that encourage discussion, identification, and mitigation of study biases, privacy concerns, and potential for harm. Here, we briefly discuss examples of approaches that may be used to promote transparency and inclusivity in these study design elements.

Models trained on biased data can perpetuate clinical biases in generated content[14]. All available details regarding data distribution of training and evaluation datasets should be reported, including patient sociodemographic information, any data imbalance, the time period when the data was collected, and any changes to best practice medical guidelines during this time period[15]. When possible, analysis of model performance across diverse patient subgroups is strongly encouraged to identify biases in downstream deployment and impact on patient care and decision-making[8,9]. This is particularly critical if training or evaluation datasets are not reflective of real-world patient diversity or clinical workflows, and external validation to assess model fairness and robustness should be performed across different data distributions if possible. For assessment of cultural and social biases, researchers should consider engaging with a diverse set of clinical evaluators. Potential clinical impacts of generative models should also be identified or if possible, assessed in real-world settings with patient-centered approaches that are inclusive of diverse cultural and social communities[16,17].

Due to the rapid development of generative modeling approaches, data privacy and security vulnerabilities also remain a significant concern[3]. Model vulnerabilities should be assessed based on up-to-date literature on privacy and security[18,19], and care must be taken to ensure that sensitive data or model outputs from sensitive data are maintained in secure environments[20]. This section provides only a brief description of potential approaches to analyzing and addressing model safety, fairness, and reliability, and we point researchers towards more comprehensive guidelines on each of these topics[18,21–23].

7.2 A new train-test split for prompt development and few-shot learning

To minimize data leakage and prevent overfitting, study design should also ensure that training, validation, and test datasets are independent of each other. While traditional machine learning methods typically rely on large, well-annotated datasets for training, newer generative models have been shown to be capable of performing tasks with minimal examples (few-shot), or even without any specific examples (zero-shot).

For simple supervised machine learning models, common train-test splits typically use about 70-80% of the data for training, ~10-15% for hyperparameter tuning, and the remainder used only for final model evaluation. For few- or zero-shot approaches, the "training" dataset can be kept to a minimal fraction of the data, still independent from the validation or test datasets. Data splits should be performed at the patient level, with all data from each patient only included in one of the splits to maintain independence. We also emphasize the use of an independent "prompt validation" dataset for prompt engineering, which should be thought of as a hyperparameter that can overfit to a dataset. Previous studies have used 5% of the data or a minimum of 50 to 100 samples[24,25] for prompt validation. While the same validation dataset should be used for prompt engineering between different models, the best prompt selected for each model may vary. For classification tasks where potential labels are provided in the prompt, the order of these labels should be randomly shuffled since models may be sensitive to the position of values in the prompt[26,27]. All prompts should be shared verbatim, along with representative model outputs when feasible, and a discussion of robustness of the model relative to specific prompts[28].

As prompt engineering is a rapidly evolving field, this checklist does not specify how to approach prompt development beyond the use of independent prompt validation datasets and appropriate randomization. We direct readers to follow best practice guidelines laid out by each model developer, which often emphasize using clear, descriptive, concise instructions, providing a value to output if the task is not applicable, and using leading cues to direct the formats of outputs[29–31]. New approaches, such as chain-of-thought approaches for reasoning tasks[32], self-consistency with shuffling[33], or training vector representations as "soft prompts"[34], should be considered when developing prompts.

7.3 Updates to baseline selection

Due to the zero-shot nature and variety of potential outputs generated by LLMs, appropriate baselines should be selected rigorously. For model baselines, both generative and non-generative approaches should be considered, particularly if the outcome is discrete and the task can be performed by non-generative models. Any post-processing of generative model outputs should be detailed in the methods, including how errors or unexpected outputs are handled. If non-generative models are used, which require training or finetuning, it's important to report their performance across various volumes of training data. Discussion of the tradeoffs between compute and cost requirements is encouraged. This allows an understanding of the scalability and efficiency of these non-generative models compared to their generative counterparts[24,35].

Given the rapid pace of model development, the most recent model available should be preferred for testing. Previous versions can serve as baselines where appropriate. Open source baselines are strongly encouraged and researchers should consider evaluating models of different sizes if available. The training datasets, context lengths, and all other model details should be reported or clearly referenced to describe their potential impact on the task being tested. If no comparable models are available, human evaluation can be used, which we detail in part 4B.

7.4 Model evaluation

Evaluation metrics for generative models should distinguish between metrics that measure overlap accuracy, which measures proportions of overlapping subunits (eg. tokens, pixels), semantic accuracy, which compare the meanings of outputs and labels, and clinical utility, which measure how models affect clinical workflows or downstream patient outcomes[36–38]. We identify best-practices for both automated and clinical expert evaluations, with a focus on metrics developed to handle the complex, unstructured outputs from generative models.

7.4.1 Automated model evaluation

Similar to traditional machine learning classification setups, accuracy, F1 scores (for imbalanced datasets), or other suitable metrics should be reported, along with class distribution, for categorical labels. For continuous outputs, such as time saved or changes to patient activity scores, which are common for assessing clinical utility of models, best practice statistical approaches and reporting should be applied, including appropriate adjustment for confounding variables and multiple hypothesis testing.

For unstructured text outputs, automated overlap scoring methods like BLEU and ROUGE are commonly used, but these only capture how well tokens match between model predictions and a ground truth reference. These provide an estimate of how well the models produce text that look correct, but do not assess whether the answers are clinically accurate, so are often poorly correlated with human evaluation on biomedical tasks[16,39]. These methods also often fail in cases of negation[40], where the model produces values such as “correct” that can match a significant proportion of the negated value “not correct” but has the opposite meaning. Additionally, these methods may not be appropriate for certain clinical tasks where reference documents typically do not exist, such as in document summarization.

Semantic scoring methods, such as BERT-based scoring methods[41] or a panel of multiple metrics[42,43], can provide more reliable evaluation, but should also be compared to human evaluation where possible to demonstrate its accuracy on new tasks. The same caution should be applied if using another AI model for semantic scoring. Although initial studies using these methods for evaluation on general, non-medical tasks appear promising[44–46], rigorous evaluation is required before applying these approaches at scale on new, clinical tasks[47].

7.4.2 Human model evaluation

Human model evaluation remains the gold-standard for assessing semantic accuracy and clinical utility of generative models. As much as possible, evaluation should be conducted in a blinded fashion, with Turing-like assessments against ground truth values or across multiple metrics to gauge the accuracy, appropriateness, bias, and other aspects of model performance[48,49]. For complex outputs or simulated scenarios, Objective Structured Clinical Examination (OSCE) type evaluations can be considered that assess model performance across multiple axes that better reflect real-world clinical encounters or workflows[16,50]. Although evaluations are dependent on the question being asked, we emphasize the need for multiple clinical reviewers and transparent reporting of inter-reviewer variability and formal evaluation guidelines.

7.5 Interpretability of generative models

Interpretability research for generative models remains an active field of investigation, and we maintain suggestions from the original MI-CLAIM checklist to apply best-practice interpretability methods when possible. These may include local interpretability techniques like LIME[51] and SHAP[52], gradient and attention analysis[53,54] for attributing importance scores to different input segments, probing methods to identify encoded knowledge[55], rule-based methods to explain model predictions as if-then-else rules[56], and counterfactual analysis to compare minimal example pairs for which language models exhibit different behavior[57]. Careful evaluation of these methods should be performed when applied to new clinical tasks[58]. Recently, methods like chain-of-thought have become popular for generating explanations to improve language model reasoning[32]. However, these generated explanations may not always align with model outputs and should not be used as a method of model interpretability[55,59].

Error analysis and sensitivity analysis (ablation tests) are also strongly encouraged as methods to better understand model behavior, particularly if evaluation datasets or models are not made publicly available. It is becoming increasingly important to understand how generative models may fail in clinical settings, which can provide insights into their capabilities and limitations beyond accuracy metrics.

7.6 End-to-end pipeline replication

Reproducible methods for generative modeling research should allow the community to replicate 1) data collection and cohort selection, 2) model development, inference, and/or deployment, and 3) end-to-end evaluation. Best-practice methods for reproducible cohort selection are discussed in Part 1 alongside study design. For reproducible model development or usage, random seeds and other hyperparameters should be reported, along with detailed descriptions of model inputs and implementation frameworks, especially if code and/or data are not provided. Due to the rapid development of generative models, accurate reporting of model versions is also crucial. As mentioned in section 1, model cards that detail model capabilities, intended use, training data and limitations, potential biases, and model risks should be provided if releasing a new model[21].

If possible, a sample of the raw data, synthetic data, or the data structure derived following patient selection as well as the processed data should be provided[6]. Use of any synthetic data and strategies for generation should follow individual journal guidelines on data reporting. Along with training data, we also emphasize the importance of releasing prompts or other in-context learning data, as well as annotation guidelines and details on metrics used for evaluation. Ideally, prompts that did not work well and corresponding results should also be reported. Additionally, to promote translatability, we encourage researchers to include infrastructure and compute requirements needed to run or develop the model as part of their methods. These may include, but are not limited to, the type and quantity of hardware used, actual or estimated costs of inference or training, and training time if applicable.

7.7 Conclusions

There is enormous potential for generative models to unlock new research directions and applications, but robust study design and evaluations are crucial for developing reproducible, transparent, safe, and diverse models for clinical research and deployment. While the focus and examples provided here pertain primarily to generative language modeling, these principles can be applied to research using biomedical vision, speech, and multimodal models as well. The updated MI-CLAIM checklist can be found at https://github.com/BMiao10/MI-CLAIM-2024. We welcome continuous community feedback as the generative modeling landscape evolves. Since best practices for generative modeling are likely to change as new research emerges in prompt engineering, model bias evaluation, and interpretability approaches, the updates presented here focus on broad differences in generative modeling compared to traditional AI model development. The updated checklist aims to formalize these guidelines for generative modeling study design, baseline model development, generative language model evaluation of model accuracy, bias, and fairness, interpretability, and end-to-end reproducibility for clinical applications.

7.8 Tables

Table 7.1 *Updated MI-CLAIM checklist for generative AI clinical studies.*

Before paper submission		
Study design (Part 1)	**Page number**	**Notes if not completed**
The clinical problem in which the model will be employed is clearly detailed in the paper.		
The research question is clearly stated.		
All cohort selection criteria and study design are detailed using precise, unambiguous language.		
The characteristics of the cohorts are detailed in the text and are shown to be representative of real-world clinical settings.		
Details on how labels were generated are described, including any annotation guidelines, level of experience of annotators, inter-annotator scores, etc.		
Which step(s) have been taken to understand model biases, privacy and security concerns, and other potential harm?	☐ Discussion ☐ Identification ☐ Mitigation	
Data and optimization (Part 2)	**Page number**	**Notes if not completed**
The origin of the data is described and the original format is detailed in the paper.		
All data preprocessing for model training or inference is described, including appropriate randomization and other transformations.		
The independence between training, validation (including prompt evaluation), and test sets has been		

proven in the paper, and data is split at the patient level.		
Details on the models that were evaluated and the code developed to select the best model are provided, including any prompt development or evaluation techniques.		
Details on post-processing for model outputs should be detailed.		
Is the output data type categorical, continuous, or unstructured?	☐ Categorical ☐ Continuous ☐ Unstructured	
Model performance and evaluation (Parts 3-4)	**Page number**	**Notes if not completed**
The state-of-the-art solution used as a baseline for comparison has been identified and detailed. Both generative and non-generative approaches are considered.		
The performance comparison between the baseline and the proposed model is presented with the appropriate statistical significance.		
Identify which type(s) of evaluations were performed, and provide clear justifications for the primary metrics used for each evaluation.	☐Overlap accuracy ☐Semantic accuracy ☐Clinical utility	
If applicable, details on human evaluation are described, including any evaluation guidelines, level of experience of evaluators, inter-reviewer scores, etc.		
Model examination (Part 5)	**Page number**	**Notes if not completed**
Relevant interpretability techniques, error analysis, and/or other approaches are applied to understand factors contributing to model behavior.		

A discussion and/or assessment of the reliability and robustness of the model as the underlying data distribution shifts is included.		
Reproducibility (Part 6)	**Page number**	**Notes**
Data transparency: choose appropriate tier of transparency		
Tier 1: complete sharing of the code and data		
Tier 2A: complete sharing of the code with synthetic data provided		
Tier 2B: complete sharing of the code		
Tier 3: no sharing of code or data		
Model transparency		
Model hyperparameters, along with infrastructure and compute requirements for running or developing the model are included, specifying hardware type, costs, and training time where applicable.		
If applicable: Model cards detailing capabilities, intended use, training data, limitations, potential biases, and risks are provided.		

References

1. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).

2. Gao, Y. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. Preprintat https://doi.org/10.48550/arXiv.2312.10997 (2024).

3. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. (2023).

4. Shah, N. H. *et al.* A Nationwide Network of Health AI Assurance Laboratories. *JAMA* **331**, 245–249 (2024).

5. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).

6. Sushil, M. e*t al*. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 1(4):AIdbp2300110 (2024).

7. Carlini, N. *et al.* Extracting Training Data from Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2301.13188 (2023).

8. Balloccu, S., Schmidtová, P., Lango, M. & Dušek, O. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. Preprint at https://doi.org/10.48550/arXiv.2402.03927 (2024).

9. Sainz, O. *et al.* NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. in *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds. Bouamor, H., Pino, J. & Bali, K.) 10776–10787 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.findings-emnlp.722.

10. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health

care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).

11. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).

12. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun. Med.* **3**, 1–8 (2023).

13. Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. (2024).

14. Zhang, H., Lu, A. X., Abdalla, M., McDermott, M. & Ghassemi, M. Hurtful words: quantifying biases in clinical contextual word embeddings. in *Proceedings of the ACM Conference on Health, Inference, and Learning* 110–120 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3368555.3384448.

15. Jones, C. *et al.* A causal perspective on dataset bias in machine learning for medical imaging. *Nat. Mach. Intell.* **6**, 138–146 (2024).

16. Tu, T. *et al.* Towards Conversational Diagnostic AI. Preprint at https://doi.org/10.48550/arXiv.2401.05654 (2024).

17. Shick, A. A. *et al.* Transparency of artificial intelligence/machine learning-enabled medical devices. *Npj Digit. Med.* **7**, 1–4 (2024).

18. Gupta, M., Akiri, C., Aryal, K., Parker, E. & Praharaj, L. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access* **11**, 80218–80245 (2023).

19. Feffer, M., Sinha, A., Lipton, Z. C. & Heidari, H. Red-Teaming for Generative AI: Silver Bullet or Security Theater? Preprint at https://doi.org/10.48550/arXiv.2401.15897 (2024).

20. van Breugel, B. & van der Schaar, M. Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data. Preprint at https://doi.org/10.48550/arXiv.2304.03722 (2023).

21. Mitchell, M. *et al.* Model Cards for Model Reporting. in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (2019). doi:10.1145/3287560.3287596.

22. Gichoya, J. W. *et al.* AI pitfalls and what not to do: mitigating bias in AI. *Br. J. Radiol.* **96**, 20230023 (2023).

23. Ning, Y. *et al.* Generative Artificial Intelligence in Healthcare: Ethical Considerations and Assessment Checklist. Preprint at https://doi.org/10.48550/arXiv.2311.02107 (2024).

24. Miao, B. Y. *et al.* Identifying Reasons for Contraceptive Switching from Real-World Data Using Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2402.03597 (2024).

25. Williams, C. Y. K., Miao, B. Y. & Butte, A. J. Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department. 2023.10.19.23297276 Preprint at https://doi.org/10.1101/2023.10.19.23297276 (2023).

26. Lu, Y., Bartolo, M., Moore, A., Riedel, S. & Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. Preprint at https://doi.org/10.48550/arXiv.2104.08786 (2022).

27. Williams, C. Y. K. *et al.* Assessing clinical acuity in the Emergency Department using the GPT-3.5 Artificial Intelligence Model. 2023.08.09.23293795 Preprint at https://doi.org/10.1101/2023.08.09.23293795 (2023).

28. Mizrahi, M. *et al.* State of What Art? A Call for Multi-Prompt LLM Evaluation. Preprint at https://doi.org/10.48550/arXiv.2401.00595 (2024).

29. Microsoft. Azure OpenAI Service - Azure OpenAI. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering (2023).

30. Liu, P. *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **55**, 195:1-195:35 (2023).

31. Liu, N. F. *et al.* Lost in the Middle: How Language Models Use Long Contexts. Preprint at https://doi.org/10.48550/arXiv.2307.03172 (2023).

32. Chu, Z. *et al.* A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. Preprint at https://doi.org/10.48550/arXiv.2309.15402 (2023).

33. Nori, H. *et al.* Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Preprint at https://doi.org/10.48550/arXiv.2311.16452 (2023).

34. Gu, J. *et al.* A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2307.12980 (2023).

35. Lehman, E. *et al.* Do We Still Need Clinical Language Models? Preprint at http://arxiv.org/abs/2302.08091 (2023).

36. Ayers, J. W., Desai, N. & Smith, D. M. Regulate Artificial Intelligence in Health Care by Prioritizing Patient Outcomes. *JAMA* (2024) doi:10.1001/jama.2024.0549.

37. Gehrmann, S., Clark, E. & Sellam, T. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. Preprint at http://arxiv.org/abs/2202.06935 (2022).

38. Goodman, K. E., Yi, P. H. & Morgan, D. J. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA* (2024) doi:10.1001/jama.2024.0555.

39. Tang, L. *et al.* Evaluating large language models on medical evidence summarization. *Npj Digit. Med.* **6**, 1–8 (2023).

40. Hossain, M. M., Anastasopoulos, A., Blanco, E. & Palmer, A. It’s not a Non-Issue: Negation as a Source of Error in Machine Translation. Preprint at https://doi.org/10.48550/arXiv.2010.05432 (2020).

41. Shor, J. *et al.* Clinical BERTScore: An Improved Measure of Automatic Speech Recognition

Performance in Clinical Settings. Preprint at https://doi.org/10.48550/arXiv.2303.05737 (2023).

42. Saidov, M., Bakalova, A., Taktasheva, E., Mikhailov, V. & Artemova, E. LUNA: A Framework for Language Understanding and Naturalness Assessment. Preprint at https://doi.org/10.48550/arXiv.2401.04522 (2024).

43. Tierney, A. A. *et al.* Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catal.* **5**, CAT.23.0404 (2024).

44. Chen, Y. & Eger, S. MENLI: Robust Evaluation Metrics from Natural Language Inference. *Trans. Assoc. Comput. Linguist.* **11**, 804–825 (2023).

45. Fu, J., Ng, S.-K., Jiang, Z. & Liu, P. GPTScore: Evaluate as You Desire. Preprint at https://doi.org/10.48550/arXiv.2302.04166 (2023).

46. Lee, H. *et al.* RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Preprint at https://doi.org/10.48550/arXiv.2309.00267 (2023).

47. Hada, R. *et al.* Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? Preprint at https://doi.org/10.48550/arXiv.2309.07462 (2024).

48. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

49. Perlis, R. H. & Fihn, S. D. Evaluating the Application of Large Language Models in Clinical Research Contexts. *JAMA Netw. Open* **6**, e2335924 (2023).

50. Mehandru, N. *et al.* Large Language Models as Agents in the Clinic. Preprint at https://doi.org/10.48550/arXiv.2309.10895 (2023).

51. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Preprint at https://doi.org/10.48550/arXiv.1602.04938 (2016).

52. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at https://doi.org/10.48550/arXiv.1705.07874 (2017).

53. Ding, S. & Koehn, P. Evaluating Saliency Methods for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2104.05824 (2021).

54. Hao, Y., Dong, L., Wei, F. & Xu, K. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proc. AAAI Conf. Artif. Intell.* **35**, 12963–12971 (2021).

55. Luo, H. & Specia, L. From Understanding to Utilization: A Survey on Explainability for Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2401.12874 (2024).

56. Sushil, M., Suster, S. & Daelemans, W. Contextual explanation rules for neural clinical classifiers. in *Proceedings of the 20th Workshop on Biomedical Language Processing* (eds. Demner-Fushman, D., Cohen, K. B., Ananiadou, S. & Tsujii, J.) 202–212 (Association for Computational Linguistics, Online, 2021). doi:10.18653/v1/2021.bionlp-1.22.

57. Yin, K. & Neubig, G. Interpreting Language Models with Contrastive Explanations. Preprint at https://doi.org/10.48550/arXiv.2202.10419 (2022).

58. Burger, C., Chen, L. & Le, T. "Are Your Explanations Reliable?" Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and Adversarial Attack. in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds. Bouamor, H., Pino, J. & Bali, K.) 12831–12844 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.792.

59. Turpin, M., Michael, J., Perez, E. & Bowman, S. R. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.

Chapter 8

Conclusions

8.1 Contributions

The field of language modeling shifted significantly with the public release of ChatGPT and rapid development of methods in the emerging generative modeling landscape in the year that followed. While these models showed significant promise on publicly available biomedical benchmarks, there have been justified concerns brought up regarding data leakage and calls for new datasets and evaluation approaches that go beyond curated question answering. This thesis contributes several new evaluations, tasks, and perspectives on clinical language modeling using both black-box LLM methods and open-source models. Chapters 2 and 3 focused on digital therapeutics and interventions and shed light on these novel therapeutics available for patients and how digital health tools impact clinical workflows and patient care. Chapter 4 developed and applied methods to assess the ability of proprietary LLMs to extract information on clinical decision making from real-world notes compared to traditional natural language processing approaches. Chapter 5 extended this method beyond prescribed contraceptive switching and demonstrated the capabilities of both proprietary and open-source language models in understanding reasons for targeted treatment switching in patients with autoimmune diseases. Chapter 6 shifted towards evaluating LLMs on a new task in developing clinical decision trees from clinical guidelines in oncology. From the learnings in the previous chapters, we conclude with Chapter 7, which provided a formalized checklist for robust, transparent, and reproducible clinical language modeling research.

8.2 Future Directions

Hospitals are uniquely positioned in the language modeling landscape, with large, proprietary, specialized medical record datasets that contain rich, untapped information about human health and disease at an unprecedented scale. We are only beginning to uncover the complexity of patient care using these data but improved facilitation of interdisciplinary work across the clinical language modeling domain is required for the clinical language modeling domain to keep up with the rapid pace of generative model development. Computer scientists looking to develop, adapt, or validate computational methods often lack real-world datasets and perspective on the challenges of clinical workflows. In contrast, hospitals and clinical researchers are often not equipped with the infrastructure or incentives to produce reproducible and scalable approaches to healthcare modeling.

While new initiatives are helping to bridge the gap between these fields, these challenges continue to limit healthcare informatic approaches and much of the constantly evolving field of language modeling has yet to make its way into clinical applications. The field is moving so quickly that even some of the work that has gone into this dissertation has already become outdated in these few short years, particularly the last few months. However, adaptation and implementation of these cutting edge algorithms to future real-world clinical applications will require a critical interdisciplinary effort from both clinical and computational researchers. To conclude this thesis, I discuss briefly opportunities to pursue at the frontier of clinical language modeling, particularly their application to individual treatment prediction and the challenges in bringing these methods towards the development of personalized medicine.

With the emergent capabilities of LLMs, there is a growing emphasis on utilizing these models for more complex physician and patient-facing tasks that may involve multi-step information synthesis, use of external data sources, high-level reasoning, or even simulation of clinical text or conversations.[7,8] In these scenarios, LLMs should not be viewed as models of language, but rather as intelligent "agents" that have internal planning capabilities that allow them to perform complex, multi-step reasoning or interact with tools, databases, other agents, or external users to better respond to user requests.[8,9] While the field of agent-based modeling is not new, with real-world applications in modeling disease and training self-driving cars, the complex, domain-specific interactions between patients, physicians, and other aspects of the healthcare system have previously been difficult to simulate. Now, with the development of new LLMs and agents designed to learn how to navigate complex data landscapes, high-fidelity simulations of clinical scenarios and complex workflows are increasingly possible.

However, there remain significant challenges in critical evaluation of LLMs in these dynamic environments and their effects on patient care, and a need for new clinical agent benchmarks that go beyond curated question answering. There must also be the development of new methods to maintain patient privacy while ensuring that data can be used responsibly to drive medical innovation. Similar to standards and regulations for the autonomous driving industry, identifying robust clinical guidelines and what constitutes a successful interaction for healthcare LLM agents will be crucial towards fulfilling the long-term goals of patients, providers, and other clinical stakeholders.

The rapid development of LLMs is happening in parallel to an unprecedented growth of clinical data collected both in and out of the healthcare system. This explosion of data and computational approaches necessitate further development of the robust, scalable evaluations, building on the methods and findings presented here, to reliably use these technologies in clinical workflows and for the improvement of patient care.

Publishing Agreement

DocuSigned by:
Brenda Miao
647B773A2DC140A...
Author Signature

5/28/2024
Date