Needs-driven, utility-oriented, standards-based operationalization of artificial intelligence for clinical decision support: a framework with application to suicide prevention

Hannah A. Burkhardt

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Trevor Cohen, Chair

Andrea Hartzler

William B. Lober

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Needs-driven, utility-oriented, standards-based operationalization of artificial intelligence for clinical decision support: a framework with application to suicide prevention

Hannah A. Burkhardt

Chair of the Supervisory Committee:

Trevor Cohen

Department of Biomedical Informatics and Medical Education

While artificial intelligence (AI) technologies increasingly permeate our daily lives, the adoption and impact of AI have fallen short of expectations in healthcare. The challenges of operationalizing AI in healthcare are complex and include interaction design (e.g. poorly designed user interfaces), model formulation (e.g. algorithmic bias, limited practical utility, trustworthiness or interpretability), and workflow context (e.g. a lack of integration into existing workflows; limited actionability). Critically, AI projects must demonstrate overall utility, balancing their costs with the benefits they confer. To achieve this utility, informatics efforts are needed before, during, and after predictive model development, to mediate effective, sustainable, and interoperable AI deployment to support clinical workflows.

In this work, I investigated how human-centered design methods, needs-driven model development, utility-oriented evaluation methods, and standards-based software design can be leveraged collectively to address the unique challenges faced by healthcare AI, and achieve clinically impactful AI implementations. The two key contributions resulting from it are (1) a

generalizable framework for the needs-driven operationalization of AI to support healthcare workflows and clinical decision making, and (2) the application of this framework to conceive, implement and evaluate AI support for suicide prevention.

To apply this framework, I used human-centered design methods to assess technological support needs for Caring Contacts, an evidence-based suicide prevention intervention, revealing opportunities for AI-based cognitive support. Using neural transfer learning from publicly available social media data, I developed accurate natural language processing models for risk-based prioritization of patient messages. Through utility-oriented evaluation metrics, I demonstrated that this model has the potential to positively impact clinical practice. Incorporating this model, I devised a standards-based, reusable, interoperable, workflow-integrated information system for cognitive support of Caring Contacts. I developed blueprints for a FHIR data representation model and information system architecture, and implemented and shared an open-source software application.

Together, this work contributes towards bridging the historical implementation gap by furthering methods for the design, development, and delivery of AI-supported interventions, and by guiding future attempts to realize the potential of AI in clinical settings.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank Dr. Trevor Cohen. I could not have asked for a better mentor. Trevor has been a friend and guiding light, always kind, generous, patient, and helpful, but holding me to high standards all the same. I deeply appreciate the innumerable hours he has spent reading and editing my drafts at all hours of the day, and hope that some of his unparalleled writing skills have stuck with me. I have learned so much from him and will forever be grateful.

I would like to thank my doctoral committee members, Dr. Bill Lober, Dr. Andrea Hartzler, Dr. Kate Comtois, and Dr. Jeffrey Heer. I am fortunate to have enjoyed their thoughtful input, guidance, and support.

I am especially grateful to Dr. Comtois. Kate is an exceptional researcher and mentor, who not only trusted me with large parts of the ISACC project, but also secured funding to support me.

I am grateful to Dr. Kari Stephens and Dr. Matthew Thompson for their tireless mentorship and support throughout my years at the University of Washington.

I would like to thank the faculty, staff, and students of the Department of Biomedical Informatics and Medical Education (BIME) for fostering a welcoming and supportive environment that encourages growth and excellence. The friends I have made here are one of a kind. I have been lucky to be part of a group that cheers each other on, but also holds each other to high standards.

This work has been possible thanks to many wonderful collaborators. Thank you Megan Laine, Amanda Kerbrat, Dr. George Alexopoulos, Dr. Michael Pullmann, Dr. Derrick Hull, Dr. Patricia Areán, Xiruo Ding, Dr. Pascal Brandt, Dr. Jenney Lee, Sierramatice Karras, Paul Bugni, Ivan Cvitkovic, Amy Chen, and Justin McReynolds for your contributions to this work.

Special thanks to the members of the Clinical Informatics Research Group (CIRG), led by Dr. Bill Lober, without whom the software systems that are part of this work would not have been possible.

Thank you also to the members of Dr. Cohen's research group for always being available to listen to presentations of my work and provide valuable feedback.

My parents have my unending gratitude for their unconditional love and support. Thank you to my mother, Manuela Burkhardt, for her unwavering confidence in me, and for her kindness, thoughtfulness, and authenticity, to which I will forever aspire. Thank you to my father, Rainer Burkhardt, who always encouraged me to aim high, provided me with the mindset, tools, and skills to succeed, and supported me every step of the way.

This work was supported by the National Library of Medicine Informatics Training Grant (grant number 67-3780, T15LM007442) and by Innovation Grant "Informatics-Supported Authorship for Caring Contacts (ISACC)" from the Garvey Institute for Brain Health Solutions.

Dedication

I dedicate this work to those who have experienced suicidal thoughts and behavior, and to those who have lost loved ones to suicide.

Chapter 1. Introduction & overview

1.1 The need for AI in healthcare

The US healthcare system today is unsustainable, inequitable, and often dangerous to patients. It is the most expensive healthcare system in the world: costs as a share of GDP rose steadily over the past two decades, reaching 19% in the US in 2020, compared to less than 13% for the OECD nation ranked second (Canada) [1], while population health indicators such as life expectancy are low [2] (Figure 1.1). Since 2019, both life expectancy and healthcare expenditures have been affected by the COVID-19 epidemic across the world, and in the United States, we have seen the steepest decline in life expectancy and the greatest increase in costs. Medical error is the third leading cause of death in the US [3]. Health disparities run rampant, with racial and ethnic minorities experiencing worse outcomes than majority groups despite decades-old efforts to address such inequities [4].

To restructure healthcare and address these issues, the so-called triple aim [5] (reducing costs; improving population health; improving patient experience) has been proposed, and has since been expanded by a fourth (improving care team wellbeing) [6] and then a fifth aim (improving equity and inclusion) [7]. The idea of the learning health system is a pathway through which these aims may be achieved [8]. A learning health system is one where "science, informatics, incentives, and culture are aligned for enduring improvement and innovation; best practices are seamlessly embedded in the care process; patients and families are active participants in all elements; and new knowledge is captured as an integral by-product of the care experience" [7]. Health information technology, and data analytic methods in particular, are key components of this vision [8,9], as they provide mechanisms to learn what works best from both research and experience and update how healthcare is delivered accordingly [10]. In 2017, the Digital Health Learning Collaborative, a taskforce of the National Academies of Medicine (NAM;

formerly known as Institutes of Medicine, IOM), identified artificial intelligence (AI) as having central importance in facilitating improvements in healthcare [7].

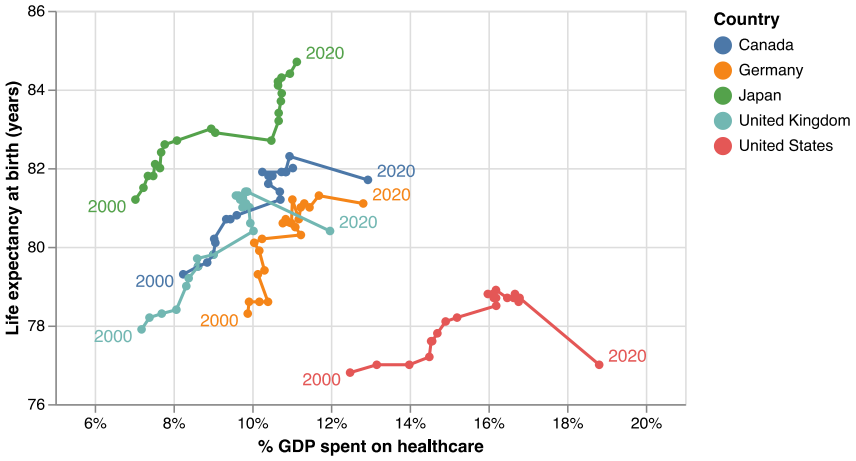

Figure 1.1 Life expectancy vs. healthcare expenditures for a selection of countries (2000-2020 OECD data).

As our lives increasingly unfold online, digital technologies and AI permeate every aspect of our lives, from the movies that are suggested to us on popular streaming services to email response suggestions and personal assistants on our smartphones. Although AI in medicine has been a subject of research for decades [11–13], the healthcare industry has lagged behind in adopting such technologies to improve medical care [7,14–17]. Recently, advances in machine learning, particularly in the area of artificial neural networks and deep learning, have resulted in performance metrics matching or exceeding those of human experts in several medical use cases [18,19], fueling renewed hope that computer-assisted technologies may soon have substantial impacts on healthcare quality and outcomes [16]. So far, such developments have not materialized, illustrating the numerous challenges in translating the power of AI into clinical care improvements. Using these techniques to make a difference in practice requires addressing a host of other concerns, ranging from ethics, policy and legal issues, costs, culture and acceptability, to integration, standardization, and interaction design considerations.

1.2 Challenges for AI in healthcare

The 2019 report by the NAM, titled "Artificial Intelligence in Medicine: The Hope, the Hype, the Promise, the Peril" [7], summarized the challenges and opportunities of using AI in healthcare settings. The authors argue that we must avoid falling victim to unrealistic expectations in order to prevent subsequent disillusionment and the abandonment of these promising technologies, as happened twice before in the so-called "AI winters": first in the 1970s, and again in the late 1980s. In outlining a path forward, the report emphasizes the thoughtful exploration and deployment of AI in a way that is safe and effective, leveraging known best practices from human-centered design, software engineering, and implementation science, and aiming for augmented intelligence rather than full automation.

AI deployments in healthcare require three properties to be successful: safety, actionability, and utility [14]. Safety means that first and foremost, patient harm must be prevented. In the augmented intelligence approach, AI provides guidance to augment clinical decision making and to help avoid errors. This "human in the loop" approach positions AI as assistance to a clinical professional, rather than an agent making autonomous healthcare decisions, addressing questions of safety; however, the interpretability and credibility of outputs are prerequisites for this approach. To be safe for all patients also means to provide benefits equitably across patient groups, such that no individual group is disadvantaged as a result of AI, or systematically excluded from reaping its benefits. Finally, the data AI models utilize must be current and correct, so that erroneous predictions may be avoided. Aspects of the information infrastructure may play a key role in this regard, as they may contribute to or hinder data fidelity [20]. Therefore, high-quality data must be available and accessible at the time it is needed as an input to a prediction model, and it must be in the expected format so it can be processed correctly.

Actionability means that healthcare AI applications must be carefully designed to meet a specific need. Demonstrations of the feasibility of high-accuracy predictive models for certain prediction tasks may motivate their integration into clinical care, only to discover that the information provided by the model has little practical value at the point of care [21]. For example, knowing an ICU patient's sepsis risk may not have any practical impact on how doctors and nurses care for the patient: Intensive care patients are likely already subject to continuous monitoring protocols, with staff ready to intervene immediately in case of an emergent crisis. Therefore, knowing that the sepsis risk is high would not be actionable for this patient. Instead of asking "How can I incorporate AI in this workflow?", researchers and developers must ask: "What do healthcare professionals need help with?" For example, it may be more useful for a clinical decision support (CDS) system to determine the most effective ICU monitoring protocol for a patient given their parameters, and issue a recommendation only if there are specific actions that could be taken to meaningfully reduce the risk.

Utility means that the deployment of the AI model must have some tangible effect on outcomes. This could be time saved by a provider, an overall reduction in healthcare costs, high-quality years of life gained, improved population health metrics, or another metric [14]. The theoretical benefits must be framed in the context of the constraints of the deployment setting, such as limited work capacity [15]. These constraints may also impact the system parameters, such as decision thresholds, that maximize clinical utility [22]. The potential utility of a system is further impacted by its financial costs, which are exacerbated by complex technical and integration requirements [20]. Healthcare organizations use a broad range of different information technology systems, with a single organization sometimes using dozens of different systems. Models and applications may already exist, but may have been developed for specific electronic health record (EHR) systems, limiting their potential for reuse. Designing sustainable information infrastructures can cut down on development and maintenance costs and enable reuse by other organizations. The consistent and systematic use of data standards and

interoperability frameworks can enable rapid application development [23] as well as the portability of informatics systems across implementation sites using different technology vendors. Additionally, making software reusable and freely available can facilitate the use of digital health interventions in resource-limited settings where development costs would otherwise be prohibitive. The principles for operationalizing AI in healthcare are summarized in Table 1.1.

Table 1.1 Principles for operationalizing AI in healthcare

Safety	Actionability	Utility
– Focus on augmented intelligence – Ensure interpretability & credibility by collaborating with clinicians – System design & data sources enable data and predictions that are current, correct, and available – Avoid algorithmic bias	– Design for demonstrated information needs – Design task-oriented model outputs – Ensure actionability by collaborating with clinicians	– Weigh anticipated system benefits against costs – Optimize parameters based on deployment constraints – Reduce development, integration, maintenance, and reuse costs by using current data standards

1.3 Making AI useful in healthcare

A considerable amount of existing biomedical and health informatics research is relevant to these principles. Human-centered design methods are being applied to health-specific problems and solutions, which has resulted in an extensive body of design principles and recommendations for a broad range of digital health interventions. Researchers have investigated AI for medicine for decades [11–13], and due to recently renewed interest by biomedical informaticists as well as computer scientists, the machine learning and predictive modeling literature is ever-growing [7,13,18,19]. The informatics community and international standards organizations continue to make leaps in developing and disseminating data standards for biomedical knowledge and processes, and provide countless freely available tools, frameworks, and resources to help develop sustainable healthcare software [24–27]. However, there is a lack of work synthesizing these advances for applications with real-world clinical

impact [14,16,17,28]. Consequently, there is a disconnect between advancing and optimizing these individual process and system components via methodological research and putting them to use in systems that work as a whole. For example, machine learning methodology researchers tend to focus on optimizing predictive accuracy, and seldom take clinical utility into account, even though high accuracy alone may not result in high utility [15].

Despite high interest in using AI to improve digital interventions and healthcare in general, there is no comprehensive framework, best practice, or guideline for the operationalization of AI in healthcare [21]. An extensive literature search revealed only one recently proposed framework [15,21]. This framework lays out steps for making AI useful in clinical practice, focusing on predictive model suitability and utility issues. However, AI has more to offer than predictive modeling (such as of risk scores), with different techniques necessitating different approaches to design, development, and implementation. For example, for systems intended to support human performance on tasks involving problem solving, decision making, and memory, informaticians may draw on cognitive engineering approaches [29]. Additionally, informatics efforts are needed before, during, and after AI components are developed and evaluated to inform overall system design and to ensure the sustainable, interoperable deployment of models into clinical care workflows. Therefore, in this work, I develop an end-to-end framework to guide health informatics projects involving AI, spanning the systematic assessment of the clinical use case and corresponding user needs (including clinical cognition and decision making), as well as design efforts involving stakeholders; development of algorithms that meet the identified needs while addressing common challenges; and implementation of sustainable, interoperable information systems with the potential to impact clinical care.

1.4 The case of Caring Contacts

Mental health is among the most critical health issues of our era. Suicide rates have increased by 28% over the past two decades and suicide is now one of the leading causes of death, with some populations, e.g. military veterans, disproportionately affected [30–32]. 1.2 million U.S. adults and 629,000 adolescents aged 12-17 attempted suicide in 2020 [31]. In 2018, the suicide rate was nearly 2.5 times the homicide rate [33] (Figure 1.2). While suicide attempts and deaths have far-reaching emotional and economic effects within the affected communities, they represent only the tip of the iceberg: 12.2 million U.S. adults and 3 million adolescents had serious thoughts of suicide in 2020 [31]. Suicidal thoughts are the result of debilitating emotional suffering, yet it can be difficult to reach out for help, and many affected individuals do not receive the support they need.

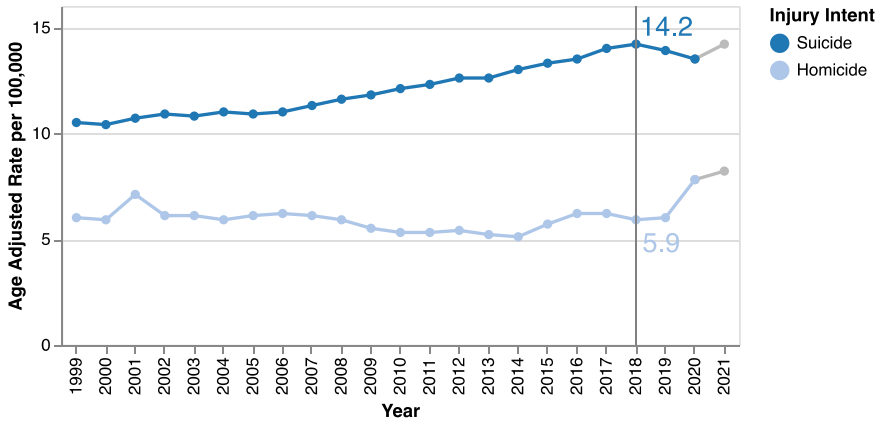

Figure 1.2 Suicide rates compared to homicide rates. Preliminary data is shown in gray. Data source: CDC.

In many cases, suicide deaths occur despite opportunities for intervention: 45% of people who die from suicide had a primary care encounter in the month leading up to their suicide [34]. Regular suicide risk screening at primary care encounters is effective in identifying individuals at elevated risk, particularly in vulnerable populations; for example, in a study of individuals with psychotic disorders, Simon et al. found that 59% of those with a suicide attempt indicated at least some level of suicidal ideation on the PHQ-9 questionnaire in the past year

[35]. The Joint Commission now prescribes several suicide prevention measures for healthcare organizations, including completing safety planning with suicidal individuals evaluated in Emergency Departments (EDs) [36], which includes determining personalized warning signs and coping strategies, identifying family members and friends to contact for both distraction and support, and listing mental health professionals and services to contact in the event of a suicidal crisis [37].

An increasing amount of evidence supports the efficacy of post-discharge follow-up contacts. Social isolation, which has been hypothesized to be a key contributor to suicide risk, is counteracted by providing patients with a sense of connectedness to members of their health care team [38,39] in interventions such as Caring Contacts [40,41]. Caring Contacts, now recommended by clinical practice guidelines [42] and used by healthcare organizations across the United States, has been shown to be effective in reducing suicidal thoughts and behaviors, suicide attempts, and suicide completion in many studies, including clinical trials [38,40,41,43]; the evidence is strongest for suicide attempts, with effects ranging from 20% to 60% reduction at 1 year after enrollment [44].

Caring Contacts entails care team members, suicide prevention professionals, or supportive staff (e.g., behavioral health providers, social workers, clerical staff) periodically sending brief messages of unconditional care and concern to individuals who are or previously were at risk of suicide (e.g. “Hope this week is going well for you.” [40]). Caring Contacts programs tend to enroll individuals with known risk, e.g. those with a suicide-related encounter. Different modes of message delivery (postal mail [39,43,45,46], emails [47], text messages [40]), schedules (e.g. monthly or every couple of months, and on special occasions such as birthdays), and levels of personalization (e.g. personalized reminders of support resources, previously mentioned coping strategies, or no personalization) have been used. The parameters of the intervention depend on the program goals, the healthcare organization’s constraints, and the specific patient population targeted. For example, Reger et al. [48] investigated the

preferences of 154 veterans regarding delivery mode, message content, and message frequency, and found that monthly messages delivered via letters or postcards were most preferred. Across three large randomized trials, Comtois et al. [40,49,50] refined Caring Contacts via text message and found that a two-way, text message-based version of the intervention was acceptable and effective in a population of military service members, while being significantly easier and cheaper to administer. In intervention designs with two-way communication, patients may reply to messages if they wish, and will receive further tailored support from Caring Contacts staff in response.

However, organizations wishing to implement Caring Contacts face challenges in deciding how to make judicious use of available resources, address patient safety, and reach recipients in a meaningful way. The Caring Contacts intervention is labor-intensive, as it requires time commitments from clinicians and support staff as well as physical and technological resources (and materials for letters or postcards in the case of analog Caring Contacts). Messages must be sent on a specified schedule, and some implementation sites customize messages for their target populations or even for each recipient. Organizations must carefully weigh the advantages and disadvantages of different modalities (e.g., mail, email, text message) for staff and resource requirements. For example, emails and texts are presumed to be cheaper and faster than sending Caring Contacts messages via postal mail, because they avoid the logistical challenges of postal mail management. It is also easier to program a sending schedule and keep track of message history with emails and texts. On the other hand, interventions based on postal mail technology are not impacted by internet availability and technology literacy.

Most Caring Contacts messages sent via mail never receive a response requiring clinical action, but emails and text messages may yield responses indicating that the recipient is experiencing distress or an acute crisis. In this case, mental health practitioners have an obligation to provide immediate support; programs must have an appropriate response and

safety plan for such cases. The need to monitor message exchanges and follow up quickly when the need arises is imperative to patient safety, necessitating a low patient-to-staff ratio in current intervention designs. As a result, Caring Contacts interventions can only enroll comparatively few patients using current formats, and many organizations have had to abandon plans to implement Caring Contacts due to overwhelming logistical and risk management difficulties with an unfavorable cost-benefit tradeoff.

At the same time, preventive healthcare is often not reimbursed by payors in the US, and clinicians are already overloaded and burned out [6]. The resulting resource shortages have precluded the broad adoption of this potentially labor-intensive intervention despite its demonstrated effectiveness. Therefore, Caring Contacts stands to benefit from computer-assisted and AI approaches that can reduce the workload and support scaling, making it possible to administer the intervention to more patients without incurring prohibitive resource requirements.

Unfortunately, it is not currently known how an AI-supported Caring Contacts application should be optimally designed, developed, and implemented. Caring Contacts provides an exemplary opportunity for the development of a generalizable framework for integrating AI into healthcare.

1.5 Aims of this work

The central research question of this work is:

How can human-centered design, value-oriented model development and evaluation methods, and standards-based software development principles be leveraged to achieve AI implementations with clinical utility?

To examine this question, I investigate the use of cognitive needs assessment methods and interoperability standards to support the development of AI models using patient-generated language data to support clinicians in the context of the Caring Contacts suicide prevention

intervention. In doing so, I follow a novel, generalizable guiding framework for the needs-driven operationalization of AI to support clinical care.

Aim 1. Establish technological support needs for an application supporting the delivery of the Caring Contacts intervention. Informatics applications must be carefully designed to meet user needs by engaging stakeholders throughout the design and development process. Stakeholders include clients, clinicians, and support staff who might interact with the application. In aim 1 of this work, I follow the principles of human-centered design to conduct key informant interviews and surveys to investigate the context of use and user needs, and then use findings to establish design considerations. Aim 1 contributes design requirements in three parts: (a) the needs around the routine administration of the Caring Contacts intervention, such as scheduling and sending messages; (b) the requirements for AI-enabled decision support components (flagging high-urgency messages and providing cognitive support for composing appropriate follow-up messages); and (c) considerations for data and workflow integration.

Aim 2. Develop a predictive model of suicide risk from patient-generated natural language and a metric for evaluating the clinical utility of such models. The analysis of patient-generated natural language offers opportunities to realize the translational potential of AI in mental health care while addressing challenges of data fidelity and availability. In aim 2, based on the requirements from aim 1, I develop neural network-based predictive models of suicide risk (used as an indicator of message urgency). These models assign scores representing the probability that a patient requires immediate intervention, and address the challenge of small clinical dataset size by exploring the use of transfer learning with publicly available non-clinical datasets. The models are evaluated not only in terms of traditional performance metrics but also with respect to their clinical utility. For this purpose, I devised a novel metric of clinical utility (average time to response in urgent messages) for triage tasks. Aim 2 contributes a risk prediction model and the novel metric.

Aim 3. Conceptualize and implement a FHIR-based software application for reusable, interoperable, workflow-integrated, AI-supported Caring Contacts. In aim 3 of this work, based on the specification established in aim 1, I develop a Fast Healthcare Interoperability Resources (FHIR)-based data representation model for the Caring Contacts intervention and operationalize it in a SMART-on-FHIR-based software application. The system ingests patient-generated text messages in real-time and leverages predictive modeling for prioritizing messages in need of immediate intervention, as developed in aim 2. The use of current health data standards and frameworks supports interoperability and integration into clinical workflows. I assess data model requirements and synthesize a FHIR representation model to meet these requirements. Additionally, I demonstrate an architectural pattern for information systems incorporating patient-generated data (PGD) at the point of care. Further, I present a discussion of the application design, functionality, and utility in light of the design considerations and requirements, considering the strengths and limitations of the system design and architecture. Aim 3 contributes a formal specification of how to use FHIR to model Caring Contacts data and workflows, an application architecture for AI-based CDS using continuously generated PGD, and an open-source, freely available software embodying these principles.

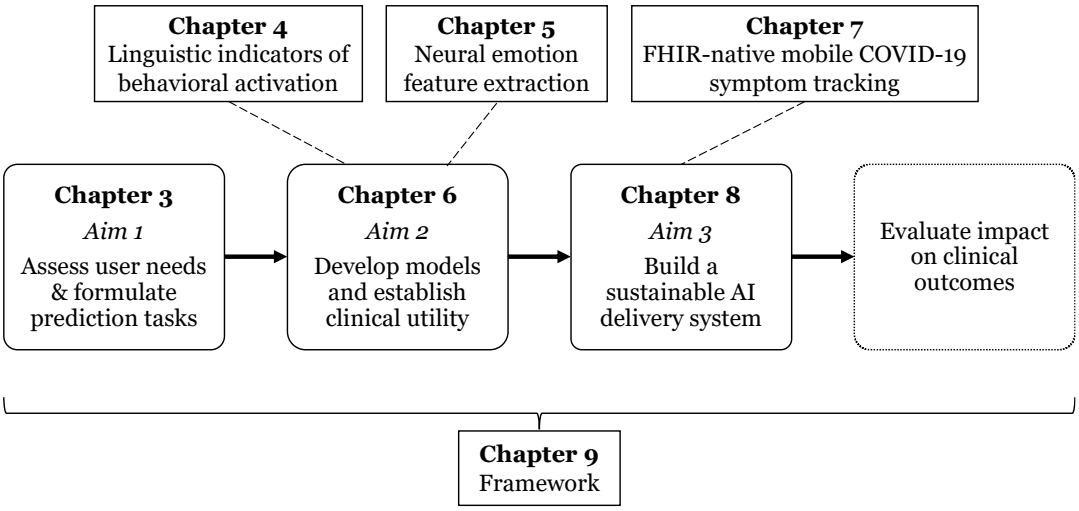

Figure 1.3 Document overview.

Finally, the findings and implications of these aims are unified in Chapter 9, which describes a generalizable framework for the needs-driven operationalization of AI in healthcare and discusses how this work embodies the principles of this framework.

Additionally, Chapter 4, Chapter 5, and Chapter 7 describe related scientific work I conducted leading up to and during this doctoral research.

In Chapter 4, I describe work in which I developed and evaluated linguistic markers of behavioral activation. The behavioral theory of depression purports that behavioral activation, i.e. the participation in meaningful, rewarding activities, is inversely associated with depression symptoms in a positive feedback loop; thus, the extent of activation, or lack thereof, is not only indicative of depression, but also serves as a therapeutic target. Harnessing distributional semantics to develop groups of related terms collectively describing the theoretical constructs of behavioral activation, I quantified behavioral activation in patient messages sent as part of text-based therapy sessions. I demonstrated that patient language can be used to measure longitudinal patient trajectories and inform intervention strategies. The techniques described in this work are relevant to suicide prevention. Linguistic indicators of suicide-related concepts, such as social isolation, could be used to extract clinically actionable insights from patient messages and inform personalized avenues to support individuals experiencing suicidal thoughts and behaviors.

In the research covered in Chapter 5, I investigated GoEmotions [51], a neural approach to extracting emotions from patient language and compared it to a well-established word counting approach. Depression affects individuals in different ways, as it is experienced as a heterogeneous combination of behavioral and thought patterns. The specifics of each patient's experience, including their emotional state, should therefore guide interventions. Fine-grained emotions may be more predictive for this purpose than sentiment, i.e. positive or negative polarity [52–54]. I showed that several extracted emotion features, including pride and disgust, correlate with depression and anxiety symptoms. These features may inform clinical

intervention on account of these emotions' relationships with established clinical constructs such as self-image and perceptions of social desirability. This work could therefore inform clinical decision support tools in the context of text-based interventions for depression therapy and related goals such as suicide prevention. The term "clinical decision support tool" is used in this work to refer to any tool that supports clinical decision making.

In a project describe in Chapter 7, I developed, described, and shared a FHIR-based, consumer-oriented application for COVID-19 symptom tracking. The COVID-19 pandemic, which took hold in the U.S. in February 2020 as I was starting my doctoral research, had a significant impact not only on population health but also on informatics research. Amidst comprehensive efforts to curb infections, public health authorities recommended that individuals self-monitor symptoms potentially indicative of an infection, such as fever, a sore throat, and shortness of breath. To support these efforts, I collaborated with public health researchers to develop and provide the general public with a mobile patient-reported outcomes (PRO)-application called StayHome. For rapid development and out-of-the-box interoperability with public health reporting agencies, we leveraged FHIR to its full extent, using it as the primary data model as well as the driver of business logic in a novel application design pattern we termed "FHIR-native". This work informed the FHIR-focused design of the suicide prevention tool developed as part of this dissertation. Additionally, both are open-source projects, and due to overlap in several areas, e.g. the use of the CarePlan resource, I was able to reuse some of StayHome's application code for the Caring Contacts application.

Chapter 2. Background and related work

2.1 Human-centered AI

2.1.1 Emulation vs. application

Artificial intelligence (AI) research has historically had two complementary but contrasting foci. They have been called AI and intelligence augmentation (IA), or, as Shneiderman calls them, the emulation goal vs. the application goal of AI [55,56]. Emulation research concerns the development of human-like devices and products. Although these may have applications, such as humanoid robots intended to socialize with elders and help them around the house, the research often focuses on goals such as passing the Turing test, rather than solving any particular application problem. On the other hand, application goal research concerns the development of solutions for human problems; these are often not emulation-based, as human qualities may be unnecessary or even a hindrance when addressing real-world problems. Artificial intelligence (AI) research has historically had two complementary but contrasting focuses. They have been called AI and intelligence augmentation (IA), or, as the influential huma-computer interaction researcher Ben Shneiderman calls them, the “emulation goal” vs. the “application goal” of AI [55,56]. Emulation research concerns the development of human-like devices and products. Although these may have applications, such as humanoid robots intended to socialize with elders and help them around the house, the research often focuses on goals such as passing the Turing test, rather than solving any particular application problem. On the other hand, application goal research concerns the development of solutions for human problems; these are often not emulation-based, as human qualities may be unnecessary or even a hindrance when addressing real-world problems.

Shneiderman argues that today, the emulation goal manifests in intelligent agents, simulated teammates, autonomous systems, and humanoid designs [55]. In his evaluation,

technologies designed to automatically complete tasks that are conventionally done by humans are misguided.

However, there is precedent for paradigm shifts in the level of automation that is considered acceptable. For example, until the 1950s, elevators were operated by people; today, all elevators are "autonomous", i.e. they are not directly supervised by human beings, and human intervention is needed only in cases of failure [57]. Many tasks could feasibly be completely automated if technology were sufficiently mature to be universally trusted.

A similar paradigm shift occurred recently in the realm of natural language processing. Shneiderman contends that a human-like understanding of natural language or images falls squarely within the emulation category. High-quality voice assistants are commonplace now; yet Amazon's Alexa, the first widely adopted in-home voice-controlled personal assistant, was introduced to the consumer market only 7 years ago, in 2015 [58]. At first, there was significant concern over such voice assistants, particularly with respect to security and privacy [59]. However, as this technology has matured, it has become ubiquitous and well-accepted. Despite Shneiderman's criticism of natural language processing and voice recognition, he acknowledges that personal assistants would not have been possible without it. Emulation research, even if it does not seem immediately applicable to any real-world problem, will therefore be needed to drive innovation.

Such paradigm shifts must be expected to occur at a different rate in healthcare, a highly regulated, risk-averse environment that tends to resist disruption. The technology adoption lifecycle, defined as part of Diffusion of Innovation theory by Rogers [60], is initially driven by innovators and early adopters, until the technology matures and is widely adopted by the majority of users, who need to see evidence that the innovation works, including demonstrated success by others who have adopted the innovation. In healthcare, cost-benefit tradeoffs must be carefully weighed with any new technology, making evidence that an innovation works indispensable. Individuals and organizations are less likely to be early adopters who do not need

to be convinced of the benefits of change. In this environment, a focus on applications, as Shneiderman advocates, is paramount. Technology that is not user-friendly or without clear benefits may not be adopted; similarly, technology that requires other systemic shifts before it will be useful may not be adopted. Instead, technology must have clear benefits, and must be demonstrably effective in delivering on its promises. Designers of any technology in the healthcare realm must be sensitive to this.

Combined with the challenges described in Section 1.2, it is becoming increasingly clear that the adoption of healthcare AI will require an increased focus on problem-oriented applications. Shneiderman's recommendation to conceptualize AI as the basis for powerful tools operating under human supervision that ultimately enable human agency is in line with the ideas of researchers in biomedical informatics, such as the authors of the NAM report [7], who increasingly recognize the need for a focus on augmentation as opposed to automation.

2.1.2 Automation vs. augmentation

The NAM recommends that the AI opportunities within healthcare are tackled via human-centered AI tools focusing on augmented AI, i.e. tools that support human beings as decision makers rather than replace them [7]. AI algorithms and human beings have different strengths: algorithms can detect subtle statistical patterns and use them to make inferences in new scenarios. Human beings have uniquely complex cognitive abilities, such as recognizing abstract high-level patterns, the ability to hold and act on moral beliefs, and empathy. Human beings assisted by technology, sometimes called human-AI teams or human-machine collaborations, combine these strengths in order to accomplish tasks better than either part of the team could hope to achieve individually [61]; for example, by a division of labor where AI automatically completes routine tasks, so that human experts may focus on more complex tasks. By extension, there is learning involved, as with any situation where human beings learn to use tools for a particular purpose; for example, human beings might learn what kinds of mistakes

the algorithm makes, and therefore when predictions are likely to be more or less reliable. This idea is central to biomedical informatics, articulated by Charles Friedman with the Fundamental Theorem of Informatics [62]: "A person working in partnership with an information resource in 'better' than the same person unassisted." In other words, informatics is concerned with the unit formed by human beings and assistive technologies, rather than technological capabilities alone; an information resource that does not provide value does not satisfy the theorem.

Importantly, technology should have appropriate safeguards in place. It should not act completely autonomously when not appropriate, and in any case, human operators should be able to take over control [55]. Human operators should be able to choose not to act on a prediction, e.g. if the prediction is believed to be incorrect or if the action is otherwise felt to be inappropriate. This means that in most cases, decisions should not be executed automatically, should be easily reversed, or contingencies should be in place in case of failure. Generally, designs should build confidence and trust, for example by including appropriate explanations [63].

In informatics, there is significant precedent for using human-centered design methods to effectively design digital health tools including clinical decision support tools. For example, Hartzler et al. identified the information needs of patients preparing to have discussions about lung cancer transplantation with clinicians, and designed and evaluated an information display accordingly [64]. Faiola et al. used human-centered design methods to design visualizations intended to support clinical decision making in the ICU by relieving cognitive load, demonstrating improved speed in decision making [65]. The amount of human-centered design research in medicine has increased exponentially in recent years (Figure 2.1). There are many examples of work applying human-centered design methods to effectively design digital health tools, including clinical decision support tools; for instance, Hartzler et al. identified the information needs of physicians preparing to have discussions about lung cancer transplantation with patients, and designed and evaluated an information display accordingly

[64]. As another example, Faiola et al. used human-centered design methods to design visualizations intended to support clinical decision making in the ICU by relieving cognitive load, demonstrating improved speed in decision making [65].

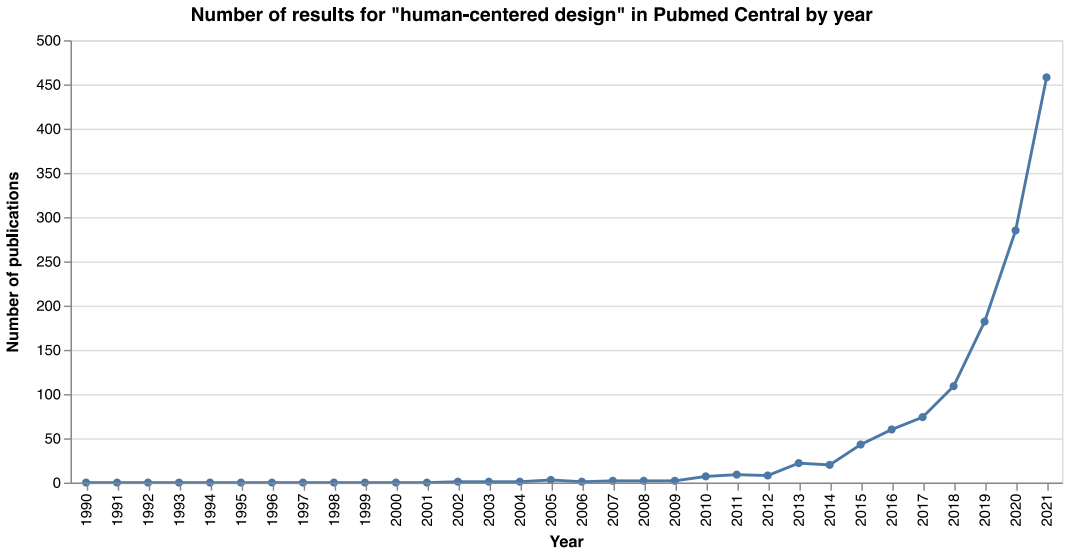

Figure 2.1 Number of results for "human-centered design" in Pubmed Central by year

However, these methods have been underutilized in conceptualizing AI use cases. A large fraction of current machine learning research is done on a few well-defined prediction tasks with benchmark datasets, i.e. datasets that happened to be available, or because certain methods are promising and important to develop further. Although the importance of standardized prediction tasks with large benchmark datasets and generalizable evaluation metrics for the advancement of machine learning methods cannot be overstated, their relevance and applicability to solutions for real problems can fall short.

For example, while Google's innovative and impactful research into detecting diabetic retinopathy from retina images [66] made an important contribution for machine learning methods research, the authors claims to clinical usefulness are contingent upon other systemic changes in how healthcare is delivered. They suggest that retinal fundus images may be used to augment or replace other markers such as lipid panels, but lipid panels are used for many purposes and therefore routinely collected. Only 14% of Americans visit an eye doctor in any

given year [67], but 78% had a wellness visit (physical or general purpose check-up) [68]. Wellness visits usually include lipid panels if indicated. Additionally, other variables give a good indication of cardiovascular risk, such as age, blood pressure, and body mass index, and assessing these variables may be sufficient to exclude cardiovascular disease as a likely diagnosis in many cases, eliminating the need to order further expensive testing such as lipid panels or retinal images. Using retina images for this purpose may be feasible if their acquisition were simplified – and there is potential for this, as evidenced by recent research demonstrating the feasibility of smartphone-based retina imaging [69]. However, as reported in 2020, Google Health researchers discovered a number of practical challenges when they operationalized their model in community clinics in Thailand, including inconsistent image quality causing high false positives, variations in workflows and conditions across clinics, concerns about the added workload imposed by the new workflow, and poor internet connectivity causing system lag and significant delays in care. The authors emphasize the importance of testing and refining how models fit into workflows, stating that "there is a need for the HCI community to develop approaches for designing and evaluating machine learning systems in clinical settings" [70]. In other words, the technology as well as the deployment and workflow design must be revised before clinical utility can be achieved.

The above example illustrates the need to carefully assess context of use and workflow needs. Recognizing that providing more information or automating certain processes does not necessarily improve overall performance, cognitive engineering is a problem-driven approach that uses users' cognitive needs to inform the design of tools such as information representation aids, advisory systems, training systems, and automation [29]. For example, designers may ask: What are the tasks? What contributes to the complexity of those tasks? What tools can facilitate this work? Human-centered approaches will accelerate AI adoption in healthcare, beginning with methods of cognitive engineering to determine challenges and opportunities in current workflows.

2.1.3 Cognitive engineering

Cognitive engineering [29,71–73], a field with roots in cognitive science, spans knowledge capture, cognitive analysis, and the development of cognitive support requirements. It is related to human-centered design, but places a special focus on human cognition. Utilizing unique methods for this purpose, cognitive engineers aim to understand and account for domain complexities; the knowledge, skills, and problem solving strategies employed by domain experts to deal with these complexities; and goals, functions, processes, and constraints specific to the work domain.

Knowledge capture methods focus on characterizing the work domain as well as the knowledge and skills of domain practitioners. Techniques include interviews, focus groups, and observation. As part of these techniques, researchers may use the critical decision method [74], cognitive task analysis [75], artifact analysis [76], and more. The critical decision method entails eliciting and analyzing past critical incidents and the processes that were involved in addressing them, ranging from assessing the situation to formulating goals and determining the sequence of actions that will accomplish these goals [72]. This includes developing an understanding of factors contributing to and hindering effective handling of these incidents. Cognitive task analysis is the analysis of individual tasks and the relevant knowledge, skills, and problem solving strategies. Developing diagrams outlining tasks, their relationships with each other, and associated challenges may be helpful. Artifact analysis is an analysis of the existing artifacts used to reduce cognitive load by offloading it to physical objects, e.g. decision aids created by practitioners as they work, or existing information technology tools, and their current role in practitioners' work, including how they support or hinder performance of work tasks. Norman argues that cognitive artifacts created by practitioners are uniquely insightful because they provide a direct window into their mental models [77]. Then, using the new knowledge of tasks and the skills, tools, and strategies used to perform them, cognitive support requirements can be developed by explicitly linking cognitive analysis results and to features in a digital solution.

Cognitive engineering has been used to better understand many types of work environments, including in the healthcare domain. For example, Bauer et al. [78] used field observations, interviews, and artifact analysis to characterize the role of paper-based flowsheets, discovering, for example, that the tabular format supported easy comparisons across time; in other words, flowsheet layout "contributes to cognitive processing in the tasks for which they are designed" [78]. At the same time, they discovered shortcomings of the paper-based format, e.g. the limited capacity of a ½" by ¾" sized box for handwritten entries and the need to manually calculate and enter fluid balance (calculated from fluid input and output). They inferred cognitive support requirements for electronic flowsheets, including the need to retain the tabular format to enable easy comparisons and automated calculation of fluid balance.

However, cognitive engineering and human-centered design methods have been underutilized for the design and development of AI. Shneiderman argues that a lack of focus on human problems and their solutions has resulted in excessive automation, and that human-centered AI will result in powerful tools that strike the balance between automation and empowering human agency where it matters [55,79]. Xu holds that human factors design is not sufficiently considered in current AI research; in his evaluation, involving human-computer interaction professionals to match AI to human needs will be pivotal in avoiding another AI winter [80]. Adler-Milstein contends that fully automating tasks such as diagnosis is unlikely to garner trust with clinicians. She argues that cognitive overload due to digitization is a major challenge for healthcare professionals today, and suggests that lightening this load by re-envisioning AI to support clinical cognition is the way forward [81]. These calls support broad application of human-centered methods to healthcare AI, with a focus on cognition.

2.2 Patient-generated natural language data for clinical decision support

Published reports of live deployments of AI in medicine largely use electronic health record (EHR) data. EHR data are a valuable resource, as they are automatically created as part

of routine care. Because they are often used for billing, they are usually comprehensive and well-structured. Additionally, EHR data are stored in the healthcare organization's own information systems, making them highly available for decision support modules running in the same system. However, there are challenges with using EHR data for predictive modeling. Models can only use data that are entered, often manually, in the medical record. The forces driving data entry, e.g. billing and regulatory requirements and clinician judgments of what is important, determine which data are entered and how and when data entry occurs. Data entry errors may happen if data are entered manually. Missing data are often not missing at random; for example, the presence of a lactate test result may in itself be informative of a patient's condition if a physician ordered the test based on suspicion of sepsis. Consequently, EHR data may be a suboptimal data source for clinical decision support (CDS) or AI models in some cases, e.g. if complete and up-to-date information is essential.

Patient-generated data (PGD) are an alternative data source for predictive modeling. PGD are data created outside the clinical care setting by patients and include health-related data such as patient-reported outcomes collected via screening surveys, physical activity data captured automatically by wearables, and more [82]. PGD also span data that are not inherently health-related but may be used for healthcare-related purposes, such as social media posts or smartphone-generated location data. Because they originate with the patient, and are often captured by consumer devices and services, PGD tend to be located in disparate systems, all external to the EHR. Sharing personal data with healthcare organizations for the purpose of improving patient care may confer improved patient engagement and patient-provider communication [83]. Using PGD for clinical decision making is facilitated by EHR integration, which is difficult for many reasons, including data interoperability, organizational infrastructure and policies, and data governance [84]. Even if data can be integrated, providers may choose not to make use of them due to a lack of directly actionable insights afforded by them [85]. Overall, PGD are underutilized in clinical care and decision making [84–86], let alone computer-assisted

CDS. This is unfortunate, as automatically captured PGD can be more accurate [87] than conventionally collected data; for example, body weight data automatically uploaded by a Wi-Fi-enabled bathroom scale would be more accurate than patient-reported body weight, which is known to be a systematically underreported quantity [88]. PGD are complete and patient-focused, in that they are recorded without being filtered through a provider lens. If they are accessible in real time, they are also uniquely suited to real-time decision support for emergent situations. Therefore, using PGD for CDS can help address the challenges of high fidelity and real-time availability.

Patient-generated natural language data as a subset of PGD further have a unique potential to inform mental health interventions. Language reflects thoughts and feelings via stylistic and word choices in addition to overt content topics [89–91]. Language is also used as the primary mode of delivery of psychotherapy, supporting its further innovative use to help providers deliver mental health care. In the context of suicide prevention, indicators of danger to self or others as well as contextual information about an emergent crisis may also be from patient language [92–94]. When used as a basis for AI, PGD – and specifically, patient language data, e.g. from text messages – can therefore illuminate mental state and intervention strategies while addressing challenges of incompleteness and recency.

Unfortunately, decision support is difficult to operationalize using patient-generated data. Facebook [95] and others [96] reportedly implemented mechanisms by which posts seriously expressing suicidal ideation are flagged directly on social media platforms; however, there have been numerous concerns with these projects, ranging from privacy concerns to whether social media companies have the appropriate expertise and resources to assess and address health concerns [97–99]. Integrating emergent signals into the clinical workflow to support decision making by medical professionals, e.g. within an established suicide prevention intervention, would address these concerns; however, integrating PGD into clinical workflows is

at an early stage [85]. At the time of this writing, there have been no attempts to integrate predictions of suicide risk based on PGD into clinical care.

2.3 Machine learning for clinical problems

2.3.1 Transfer learning

AI and machine learning applications have seen fewer successes in health and healthcare than in the consumer products realm. Machine learning, and particularly methods of natural language processing (NLP) that use deep neural networks, have an unprecedented need for data, requiring training examples in the order of magnitude of thousands to millions. In the general domain, public or private data can be harvested readily for secondary use. For example, Bidirectional Encoder Representations from Transformers (BERT), a transformer model that was first reported in 2019 by Devlin and colleagues at Google [100] and that has inspired a rich body of work focusing on large, deep language models for natural language processing, is trained on publicly available unlabeled data: a corpus of over 11,000 full-text books (800 million words), as well as the entirety of articles in Wikipedia, a publicly available encyclopedia or articles written by independent contributors (2500 million words). Commercial systems, such as the recommendation systems [101] used by Amazon and Netflix, are commonly trained using large, privately held datasets, e.g. as collected from millions of customer transactions.

In the medical domain, fewer large datasets are available. Some data are publicly available, such as post-market drug safety surveillance reports collected and archived by the Federal Drug Administration (FDA) in the FDA Adverse Event Reporting System (FAERS) [102]. Other datasets are made available to researchers formally requesting access under reasonable conditions of responsible custodianship; for example, Medical Information Mart for Intensive Care (MIMIC), a de-identified dataset of EHR data acquired during routine hospital care representing 50,000 hospital admissions, can be accessed under the terms of a data use agreement after researchers complete relevant training [103]. De-identification is particularly

difficult for text data and may be incomplete, necessitating such data use agreements to ensure researchers' good faith. Finally, researchers with ties to healthcare organizations, e.g. academic medical centers, may be able to harness EHR data collected as part of ongoing care for secondary use.

However, for many more clinical use cases that could greatly benefit from AI, there are no large publicly available datasets. Many types of data may not be routinely collected, and therefore unavailable from operational EHR systems, including patient-generated data. Additionally, data are only routinely collected for inventions that are already part of standard care. In these cases, data may be purposely collected, or there may be an opportunity to obtain data from research conducted for other purposes (e.g. clinical trials). Furthermore, creating labeled datasets presents a challenge for both primary and secondary use data. Although unlabeled data have uses, e.g. for creating domain-specific pre-trained models, many applications will require labeled data; in the medical domain, clinical experts must be consulted to create labels. Both purposeful collection and labeling cause dataset size to be limited by manpower. Finally, even if the resources required to collect and label data have been invested, many datasets cannot be shared: Health data are sensitive in nature and must be safeguarded carefully, lest we cause harm to research participants by disclosing their private information to unauthorized parties; this applies even to de-identified data, as cases of personal identities being linked to data that have been de-identified according to established criteria have been reported [104,105].

Fortunately, it may be possible to leverage large general-domain datasets and models for clinical problems using transfer learning. Transfer learning enables the application of signal from one task to another, related task. This is reminiscent of human learning, where knowledge from one learning event is applied to a different but related task; for example, my 2-year old niece learned to identify ripe tomatoes using color and other aspects of appearance, and now demonstrates remarkable accuracy in assessing the ripeness of a broad assortment of fruits and

vegetables. Recent developments in transfer learning have resulted in tremendous improvements in predictive performance across a wide range of NLP tasks, including drug efficacy classification [106], human activity recognition [107], and cross-lingual learning [108].

Contextualized pretrained language models are the perhaps most widely used transfer learning approach to NLP today. Devlin's BERT is a pretrained language model that can be repurposed for a wide range of NLP tasks and has spawned a cascade of BERT-related research, including the development of many domain-adapted BERT models, further trained on domain-specific corpora to increase specificity of the learned representations to tasks within the domain space. A field of research in its own right, domain adaption has yielded models such as BioBERT, further trained on biomedical research literature [109]; ClinicalBERT, further trained on clinical notes [110]; and MentalBERT, further trained on mental health related Reddit posts [111]. BERT learns to generate contextual representations for words or sentences using the context in which they appear. Notably, representations learned from similar domains but for completely different tasks can be effectively transferred. The original task BERT is trained for is predicting a word given its context, i.e. predicting a hidden word given the words and sentences to its left and right (bidirectional). However, BERT representations can be effectively transferred towards other tasks, such as classification, by using the pretrained weights for input and hidden layers, and changing the classification layer to modify the types of outputs that are produced. In this way, the fine-tuning phase of the transfer learning process involves learning how to combine the existing linguistic information for the purposes of solving the new prediction problem.

Transfer learning thus provides a unique opportunity to enhance clinical prediction tasks with signal derived from publicly available data. First, the proliferation of domain-adapted pre-trained language models offers opportunities to leverage large, unlabeled, domain-specific corpora. Second, publicly available labeled datasets hold promise to augment clinical data, helping to overcome its limitations. However, transfer learning for this purpose has been

underutilized. As further described in the next section, a wealth of suicide risk prediction research has been conducted using only public data (specifically, social media posts) or only clinical data (specifically, EHR data); however, to the best of my knowledge, none have leveraged transfer learning to combine them.

2.3.2 Suicide risk prediction

2.3.2.1 Structured EHR data

Suicide-related machine learning models have been developed and evaluated with various data sources, machine learning methods, and performance metrics. The availability of data collected as part of routine healthcare in the EHR for secondary use has enabled a range of predictive modeling applications, including flagging patients who show evidence of elevated sepsis risk [112], are likely to miss upcoming appointments [113], or may benefit from advanced care planning [114]. Suicide-related prediction models using EHR data have proliferated as well. For example, Simon et al. [115] used logistic regression models to predict suicide attempts and suicide deaths in a dataset of almost 3 million patients across 7 institutions, with incidence rates of 0.82% and 0.04%, respectively. Their models achieved an AUROC (C-statistic) of 0.853 for predicting attempts and 0.833 for predicting deaths in the 90 days following a primary care visit. The most important predictors for attempts included the presence of a depression diagnosis, suicide attempt history, drug or alcohol abuse history, and a high PHQ-9 item 9 score in the past year. The most important predictors for suicide death included having a mental health emergency department visit in the past 3 months or mental health inpatient stay in the past year, alcohol abuse history, benzodiazepine (an anti-anxiety medication) prescription, and presence of a depression diagnosis. In a cohort of 118,252 patients with a 0.21% incidence rate, Zheng et al. [116] used a deep neural network with 3 hidden layers to predict suicide attempts in the following 1 year using a range of structured EHR data including demographics, diagnoses, procedures, and medication orders, achieving an AUROC of 0.769. The neural model

significantly improved upon their logistic regression baseline model, which achieved an AUROC of 0.607. Although individual feature importance within the model was not assessed, the features most strongly associated with increased odds of a suicide attempt included suicide attempt history, mental health disorders, and substance abuse history. Walsh et al. developed a random forest model for suicide risk detection using structured EHR data including demographics and diagnoses, as well as medication data extracted from clinical notes, in a cohort of 3,250 patients who had an ICD code for self-injury with confirmed suicidal intent, and 12,695 controls with no history of suicide attempt [117]. The most predictive features were demographic factors (age, gender, and race), as well a history of suicide attempts and mood disorders. The model achieved an AUROC of 0.770 for predicting suicide attempts within 7 days, compared to 0.804 for the Columbia Suicide Severity Rating Scale (C-SSRS) standardized instrument; combining the model with the instrument resulted in an AUROC of 0.907 [118].

2.3.2.2 Unstructured EHR data

While structured EHR data such as diagnoses and attempt history are clearly highly relevant to predicting future attempts and deaths, clinical notes contain rich information that may not appear in any structured or coded data elements. This information may further benefit predictive models. McCoy et al. [119] conducted a survival analysis of suicide death that included both structured data such as demographics and visit history, as well as measures of positive and negative valence in discharge notes, quantified using a dictionary-based word-counting approach. The dataset contained 845,417 hospital discharges with a suicide death incidence rate of 0.1% in the follow-up period (up to 9 years, median 5.2 years). The model achieved a modestly improved AUROC (C-statistic) of 0.741, compared to 0.737 for the baseline model that used coded data only. There have also been efforts to extract clinically relevant constructs from EHR notes; for example, Zhang et al. [120] extracted suicide-related psychiatric stressors such as physical and sexual abuse, health issues, and pressure from work or school

from clinical notes via named entity recognition, demonstrating high correlations with suicidal behaviors.

2.3.2.3 Patient-generated data

Regarding PGD, a large body of work leverages social media data for NLP-based suicide classification as well as depression detection. Depression is closely related to suicide because of the increased incidence of suicidal behavior in this condition. Facebook is reported to have developed an approach to flagging user posts that are concerning with respect to suicide, integrating n-gram linear regression with metadata such as posting time [95]. In a crowdsourced dataset of 171 depressed and 305 non-depressed users with an average of 4,533 twitter posts each, De Choudhury et al. demonstrated that features extracted using Linguistic Inquiry and Word Count (LIWC) [121,122], including valence of affect and pronoun usage, as well user engagement metrics (e.g. reciprocity) and usage of depression-related terms, predicted depression with 74% precision and 63% recall [123]. LIWC is a dictionary-based word-counting tool with predefined term categories. It is widely used by researchers investigating connections between mental state and language, and has an extensive track record of validation [122]. Huang and colleagues [124] developed models for classifying posts on Chinese microblog platform Sina Weibo (similar to Twitter). The authors identified and verified 53 suicide deaths and collected 30,000 posts from the corresponding users, along with 600,000 posts from 1000 other, randomly selected users. They combined emotional valence (positive and negative) measures, pronoun counts, part-of-speech tags, and metadata such as posting time. Their best models achieved 79% precision and 60% recall. The 2019 CLPsych Shared Task competition challenged participants of the CLPsych conference [125] to develop predictive models of suicide risk in a dataset of Reddit posts annotated by experts and crowdworkers [126], resulting in several high-performing models of various types, with the best-performing model combining a support vector model with several neural network-based models in an ensemble. This best model achieved an

F1-score of 0.922 for distinguishing unconcerning posts from those indicating some level of suicide risk. In the 2021 CLPsych Shared Task challenge, participants developed machine learning models to assess suicide risk in Twitter posts of 97 individuals who died by suicide or survived a suicide attempt within 6 months, and 97 matched controls, donated through OurDataHelps.org [127]. The best model achieved an F1-score of 0.815, using a dictionary-based approach that used LIWC, priors informed by domain knowledge, and logistic regression [128]. Priors were calculated from LIWC category effect sizes reported in a previous study by Eichstaedt et al. [129], who investigated the association between depression and LIWC categories in Facebook posts.

Finally, researchers have investigated using patient-generated natural language to improve our understanding of the underpinnings of mental health disorders. A large body of work investigating the linguistic manifestations of schizophrenia and psychosis uses natural language data collected as part of research studies [130]. Such studies have furthered understanding of cognitive processes and deficits in mental disorders. For example, one study elicited speech with prompts such as “tell me the story of Cinderella” and used latent semantic analysis to show that in thought disorders, semantic coherence is disrupted [131]. Purposely elicited speech may also be used to predict future onset of mental illness: Gooding et al. predicted schizophrenia onset after 10 years on the basis of thought disorder identified via manual linguistic analysis of interview transcripts with 94% accuracy [132].

2.3.2.3.1 Clinical settings

However, specialized data collection is infeasible for the goal of leveraging real-time natural language analysis to improve ongoing patient care. Applications must use data that is readily available at the point of care, so it is more feasible to use language exchanged in the context of ongoing clinical care. Sonnenschein et al. used LIWC to analyze transcripts of routine cognitive behavioral therapy sessions conducted with 85 patients, containing over 500,000

words spoken by patients, and found that “sadness” and “anxiety” words were significantly associated with depression and anxiety diagnoses, respectively [133]. However, there are complexities associated with the face-to-face setting, e.g. the requirement to transcribe audio signals to written text before it can be analyzed. The recent spike in adoption of telehealth solutions for psychotherapy [134] has yielded data that is automatically collected in real-time. Research investigating clinical decision support tools based on such data has direct applicability to clinical settings, as the required data would be readily available to a deployed tool for real-time inference.

My work described in Chapter 4 is an example of such research. In a dataset of 10,000 patients undergoing chat-based psychotherapy while completing depression (9-item Patient Health Questionnaire, PHQ-9) and anxiety (7-item General Anxiety Disorder questionnaire, GAD-7) questionnaires at regular intervals, I developed and evaluated linguistic markers of behavioral activation [135]. Behavioral activation is a psychological construct related to planning and participating in pleasant activities, a behavioral pattern that is reduced in depression; because intentional reinforcement of such behaviors reduces symptoms, the activation of such behaviors is targeted in behavioral activation therapy [136]. Markers corresponded to depression severity scores and predicted longitudinal trajectories previously determined by latent growth analysis, i.e. whether a patient experienced improvements over time.

In another study using the same dataset, described in Chapter 5, I investigated LIWC variables with an established relationship to depression and anxiety, as well as fine-grained emotions extracted using a BERT-based neural approach, finding that they were highly explanatory of depression and anxiety symptom scores in mixed-effects linear regression. Additionally, using a random forest model, I found that the LIWC features were complementary with fine-grained emotion features for the purpose of predicting depression and anxiety status [137]. Depression and anxiety are highly heterogeneous, i.e. they are experienced in many different ways; in the context of psychotherapy, a clear understanding of patients’ emotional

experience is a prerequisite to tailoring mental health support strategies accordingly. Therefore, such features have direct applicability to real-time clinical decision support.

Sharma et al. explored the use of real-time AI insights to improve empathy in text-based peer support conversations. They developed an approach to measuring empathy in messages written in response to individuals seeking mental health support from peers on the TalkLife platform, along three dimensions (emotional reaction, interpretation, and exploration) [138]. They then used this approach to develop and evaluate a computational model based on reinforcement learning, which is capable of suggesting revisions to improve empathy, in line with best practices in therapy [139]. They conducted a pilot test deployment on the TalkLife platform, asking active users to utilize their tool to improve empathy in their peer support messages. According to human raters, the rewritten messages were more empathetic than the originals 47% of the time, equivalent 16% of the time, and worse 37% of the time [140].

In summary, there is promising evidence of the utility of patient-generated data for improving mental health care, including suicide prevention. However, work utilizing patient-generated natural language data to support ongoing care remains sparse, particularly in comparison to work utilizing non-clinical data such as social media posts.

2.3.3 Metrics of model performance

Suicide is a rare event. Measuring the performance of predictive models is complicated by the resulting class imbalances. The area under the receiver operating characteristic curve, or AUROC, is a tradeoff between sensitivity and specificity over a range of decision thresholds, and is commonly used, but may not be a good indicator of performance in tasks with inherently class-imbalanced data, such as those related to suicide [141]. Other available metrics, such as the area under the precision/recall-curve (AUPRC; a tradeoff between precision and recall over a range of decision thresholds) and F1-score (the harmonic mean between positive predictive value, or precision, and recall, at one decision threshold), may be better choices; however, they

too may not capture how useful the model will be in a real-world setting. Shah, Milstein, and Bagley [17] argue that ultimately, the only metric that matters is whether a prediction from a model results in a beneficial change in patient care. Although such beneficial changes, e.g. improvements in patient outcomes or provider satisfaction, cannot be determined with certainty at model development time due to the number of unpredictable factors influencing clinical utility, some performance metrics will approximate this utility more closely than others.

In suicide risk prediction modeling, such alternative metrics of how much utility a predictive model can provide have been reported. Motivated by the reality of resource constraints in the healthcare system, Shing et al. [142] re-conceptualized the suicide risk prediction problem as a prioritization task. A ranked retrieval problem, the task is to assign good relative priorities, i.e. sorting examples in a way that makes positive examples likely to occur near the top, and negative examples likely to occur near the bottom. For example, in an internet search, the best (i.e. most likely to be positive) results should be ranked at the top. In a setting where a user works through results in order of priority (i.e. from most to least likely to be positive), with a finite capacity to review results, the relative ordering of results determines how many positive examples are addressed and how many are missed. Time-biased gain (TBG) [143] is an information retrieval evaluation metric used in the setting where a human expert reviews positive predictions in order, investing a certain amount of time per item to verify the prediction and take appropriate action, under the constraint of a fixed time budget. Shing et al. [142] further refine this idea for the specific task of reviewing posts of social media users in a hierarchical manner: Users are ranked based on a user-level risk score; reviewing an individual user result entails reviewing that user's individual posts in order of post-level risk score. They then develop a hierarchical attention approach that jointly optimizes these ranking tasks. In the social media post review setting envisioned by the authors, the resulting model can be expected to increase the number of at-risk users that can be identified in a given amount of time, compared to a model optimized for a more traditional metric.

2.4 Designing interoperable, reusable, and scalable health information technology systems

Informatics implementations may fall short of making real-world impact despite being well-designed and meeting user needs, for reasons of utility. Financial or organizational costs of development may be prohibitive for organizations in low resource settings, and even in well-funded settings, maintenance costs may be too high to continue the informatics interventions after dedicated research funding runs out. Costs are partly driven by the need to integrate with different platforms or because available tools use proprietary formats and protocols that are difficult to reuse in different contexts. However, international standards organization Health Level Seven International (HL7) has published healthcare data standards and frameworks, e.g. Fast Healthcare Interoperability Resources (FHIR) [26] and Substitutable Medical Applications Reusable Technologies (SMART) [25], that are now well known and widely adopted. Relying on these formats and frameworks can therefore reduce the costs of implementation and maintenance, both at the initial development site and at other healthcare organizations wishing to implement similar interventions, by reducing staff requirements (less time needed to become familiar with data formats and exchange protocols) and technical resources needed (open-source tools, such as the HAPI FHIR server, are available). Further, integration with EHRs and other standards-enabled healthcare information systems is, in theory, automatically supported. For these reasons, the economic cost savings of achieving fully interoperable health information systems has been estimated at $77.8 billion [144], and the US government now mandates the implementation of FHIR by HIT organizations [145,146].

2.4.1 FHIR

FHIR is the successor to the HL7 messaging standard, which was only intended for messaging. In contrast, the FHIR specification includes not only comprehensive data exchange protocols (API specifications), but also defines complex search functions, data conversion

operations, task and workflow modeling components, and of course, an extensive list of detailed resource specifications. In some cases, FHIR resources may be sufficient to model most or all of an application's data representation needs. This allows applications to further reduce the duplication of data modeling efforts: simple applications, such as those with few functions outside of creating, reading, updating, and deleting (CRUD) data, may be modeled completely with FHIR resources and operations. As described in Chapter 7, a team of collaborators and I described this approach, termed FHIR-native, and used it to develop StayHome, a mobile-friendly, interoperable, reusable symptom tracker application for COVID-19 self-monitoring [23]. FHIR was used as the primary data structure. Due to its status as a validated and internationally accepted data format, FHIR is commonly used for exchanging data, with dedicated modules performing on-demand conversions between operational data structures and FHIR formats when data exchange APIs are invoked; however, the FHIR-native approach further extends the use of FHIR by utilizing it as its internal data structure as well. This enabled the use of HAPI - an open-source, freely available FHIR server - as the primary operational database server implementation, saving the effort of developing one from scratch. Because the FHIR-native approach relies almost completely on FHIR data structures and APIs, it maximizes time saved on data structure and API design: the initial functional version of StayHome was published and made available to the public only 2 months after development started. Because StayHome is completely open source and makes extensive use of freely available resources, myself and others will not only be able to build on the FHIR-native approach, but can reuse the software code itself.

In addition to structural interoperability, semantic interoperability is essential. FHIR is intentionally highly flexible, allowing certain concepts to be represented in different ways. Although this can be a barrier to true portability and reusability of FHIR products, it cleanly separates knowledge engineering problems from issues of technical capability. This allows team members with clinical expertise to contribute early on. However, semantic interoperability thus

requires the development, sharing, and reuse of technical documentation on how to use FHIR for a particular use case. This purpose is served by Implementation Guides (IGs) and FHIR Profiles. For example, how to use FHIR to model questionnaires, questionnaire responses, and related knowledge artifacts is well-defined in the Structured Data Capture (SDC) IG.

Although the use of FHIR for patient-generated data has been largely limited to this kind of data – i.e. patient-reported outcomes (PROs), which are usually collected with structured questionnaires – other uses have been described in Profiles. As PGD become more ubiquitous, guidance and tools have increasingly become available for this category of use cases [147]. However, many specific uses remain to be explored and are as of yet without specific guidance. Newly developed use cases for FHIR should therefore develop technical documentation, such as IGs and profiles, particularly if it is expected that many different implementations sites may benefit from this guidance.

2.4.2 SMART-on-FHIR

SMART-on-FHIR [25] is an authentication and integration framework originally conceived in 2009 (as SMART Classic) as a way to position the EHR as a platform for third-party applications. After the release of FHIR as an HL7 standard in 2013, it was re-conceptualized to use FHIR as its basis [148]. SMART-on-FHIR allows users to launch custom FHIR-based applications directly from the EHR context, with the application interface being shown as part of the EHR interface, removing the need to manage separate logins or windows. SMART-on-FHIR entails the exchange of authentication and authorization credentials as well as FHIR resources with the host system. The host system is most commonly an EHR, but could be any system that implements the FHIR specification; as a result, SMART-on-FHIR apps can easily be deployed as standalone systems by substituting a bare-bones FHIR server, such as the open-source HAPI implementation [149]. Additionally, applications are portable between FHIR-enabled systems without requiring extensive customization. Although differences in how

resources are used to represent site-specific data may require customization of business logic, the technical integration components are universal, so an application developed to work with an Epic system could be readily deployed in, for example, a Cerner system. SMART-on-FHIR apps can be written by independent app developers who have complete control over business logic and user interface, while allowing the appropriate use of FHIR resources internal to the host system. In this way, it enables completely customizable CDS functions without requiring EHR vendors to support each app, and without disrupting the user workflow.

2.5 Guiding frameworks for useful, practical AI deployments

An extensive literature search revealed only one recently proposed framework guiding the operationalization of AI in healthcare. Developed in 2020 by members of Dr. Nigam Shah's research group at Stanford, the framework [15,21] outlines sequential steps in the process of designing, developing, implementing, and evaluating AI in the clinical setting. Jung et al. [15] describe that informaticians should begin by clearly articulating the modeling problem and the intervention triggered by the model's output: What is the prediction target and what data are used to make the prediction? Given a prediction, what action would someone take? It is important to formulate the model in terms of data that are available at prediction time. Next, an appropriate model is developed and validated, ensuring the model is fair. Performance metrics should be carefully considered to determine the best candidate model. If a performant and feasible model can be developed, informaticians then proceed to design a deployment strategy and assess the resulting system's potential utility, weighing the costs and benefits in the context of constraints and workflow requirements. Given that a favorable cost-benefit tradeoff is anticipated, the system may then be deployed and prospectively evaluated in a clinical trial.

Jung et al. consider issues of utility in a novel way, positing that the overall value derived from using predictive models in practice is constrained by the clinical setting in which they are deployed. Performance metrics traditionally used for machine learning models, e.g. AUROC or

the F1 measure, may therefore not be good metrics of the model's value in practice. For example, in a ranked retrieval setting where human experts review documents in order of priority, with a fixed constraint on total time spent, metrics such as hierarchical TBG [142], which captures the number of items of interest discoverable in a ranked list given a time budget, may be more appropriate. Similarly, depending on the workflow context, false positives, false negatives, true positives, and true negatives may have distinct costs or benefits, introducing nuance into how sensitivity and specificity should be balanced to maximize utility. One example is described by Jung et al. [15], who evaluated a model recommending patients for enrollment in advanced care planning, in a clinic where the number of seats in the intervention is limited – a work capacity constraint. In this case, low specificity may be very expensive, as every false positive (a patient being enrolled in the intervention even though they are unlikely to benefit from it) fills one of a finite number of slots, removing an opportunity to enroll a "true positive" (a patient who would benefit greatly from the intervention). Bayati and colleagues [22] additionally consider parameters of the deployment environment to determine clinical utility. They describe an approach to calculating the value, in US Dollars, of deploying a model predicting readmissions, given the probability of a readmission, the cost of readmission, and the cost of an intervention that reduces the probability of readmission. If a readmission is unlikely, the cost of the intervention is higher than the cost of readmission weighed by its probability, and is thus not justified. Conversely, there is a threshold probability of readmission at which the intervention is expected to have a net cost benefit. The model's goal is to determine each patient's probability such that the most beneficial course of action can be selected. The probability threshold at which the model must discriminate well is thus directly determined by the cost of readmission and the cost of the intervention.

Both examples support Li et al.'s [21] framing of clinical utility from a systems perspective. They argue that AI-supported informatics technology should be designed holistically, starting with a clinical problem, and ending with a complete system that can be

evaluated in terms of how well it solves that clinical problem. Human-centered design methods and user experience design, including stakeholder interviews and workflow analysis, should precede design and development efforts. AI components, though core enablers of the resulting software system, must be designed, developed, implemented, and evaluated in the context of the larger software system.

Jung's framework focuses on the important issues of algorithm suitability and utility, specifically in the context of predictive models such as risk scorers. While innovative in its considerations of the constraints of the system in which an algorithm would be deployed – e.g. data availability at prediction time and limitations on clinical utility imposed by work capacity – further informatics efforts are needed before, during, and after the model development and evaluation steps outlined in this framework. First, systematic assessments of clinical and information needs are needed to inform AI components. Such assessments may reveal that AI beyond predictive modeling is needed, e.g. for cognitive support helping users decide on follow-up actions after an elevated risk has been identified. Such AI would then require dedicated design, development, and evaluation efforts, e.g. to determine how best to organize information. Second, model evaluation metrics that effectively capture the magnitude of the benefit the system is expected to provide may need to be developed. Third, to support ongoing efforts to promote interoperability and reusability, developing sustainable, standards-based information models and software systems should also be a key concern; such considerations may affect calculations of utility, and must therefore be incorporated as a key part of any operationalization project.

2.6 Contributions

In light of the gaps in the current literature, this work makes the following contributions.

Biomedical informatics contribution. AI approaches must be carefully tailored to meet the needs of specific user-defined tasks, and must be evaluated in terms of their utility for

those tasks. Thus, general machine learning approaches often fall short. This disconnect is a monumental problem in the current landscape of healthcare AI. The overarching contribution of this work is a generalizable framework for the needs-driven operationalization of AI to support workflows and clinical decision making in healthcare. Compared to prior guidance, it highlights the need to comprehensively assess and design for users' cognitive needs; to develop and make use of clinically relevant metrics of model performance; and to design and develop standards-based software. This work therefore contributes to understanding and resolving this disconnect.

Additionally, there is little precedent for NLP-based suicide risk assessment using clinician-patient communications to inform clinical practice. Aim 2 contributes the knowledge that neural network-based methods can perform well in clinical settings with data scarcity, such as suicide risk classification in patient communications, by leveraging large, publicly available datasets, e.g. of social media posts, via transfer learning. Aim 2 further contributes a novel metric of a model's clinical utility that incorporates constraints of the clinical environment.

Human-centered design contribution. Aim 1 elucidates important insights into the cognitive demands of administering the Caring Contacts intervention and contributes design considerations for digital tools supporting healthcare professionals in meeting these demands, including considerations for AI-based cognitive support tools. These findings may be generalizable to other mental health informatics use cases.

Health data standards contribution. Aim 3 of this work contributes a roadmap for using FHIR to represent Caring Contacts artifacts and workflows, including blueprints for a FHIR data representation model incorporating patient-generated natural language data, and an application architecture for applications continuously ingesting externally generated text data, processing them with machine learning, and integrating them into EHR-based workflows for CDS. Aim 3 also contributes open-source, freely available, reusable code that embodies and exemplifies the use of FHIR and SMART-on-FHIR to enable interoperability, portability, and reusability of AI-based decision support tools. This ready-to-use software artifact has broad

applicability and the potential to empower healthcare organizations to improve patient outcomes.

Behavioral health contribution. Caring Contacts is an intervention with demonstrated effectiveness [44], but logistical challenges and low capacity have hindered its adoption. By developing a novel information system that leverages AI to lighten the workload burden of the intervention, I address these challenges. Data standards and interoperability technologies enable the reuse of this information system across healthcare institutions and further tips the cost-benefit balance for potential implementation sites discouraged by the intervention's requirements. This work contributes an open-source software artifact that can be directly reused by other healthcare institutions, potentially improving the adoption of this effective but under-utilized intervention, and benefitting numerous individuals at risk of suicide. At the time of this writing, plans are underway to further refine and deploy this tool across a range of settings including clinics and service organizations serving veterans.

Chapter 3. Identifying opportunities for informatics-supported suicide prevention: the case of Caring Contacts

In the work described in this chapter, I began the process of designing, developing, and implementing an informatics tool for Caring Contacts by establishing technological support needs, engaging stakeholders in a formal needs assessment. I applied ideas of cognitive engineering to better understand the clinical problems in suicide prevention, the cognitive tasks involved in solving them, and the strategies human experts employ in this process. Based on my findings, I developed design considerations, which informed the subsequent aims of this work.

A version of this chapter was previously published by the American Association for Medical Informatics (AMIA) as an open-access article.

Burkhardt HA, Laine M, Kerbrat A, Cohen T, Comtois KA, Hartzler A. Identifying opportunities for informatics-supported suicide prevention: the case of Caring Contacts. In: *AMIA Annu Symp Proc* 2022.

This work was recognized with AMIA's Distinguished Paper Award.

Abstract. Suicide is the tenth leading cause of death in the United States. Caring Contacts is a suicide prevention intervention involving care teams sending brief messages expressing unconditional care to patients at risk of suicide. Despite solid evidence for its effectiveness, Caring Contacts has not been broadly adopted by healthcare organizations. Technology has the potential to facilitate Caring Contacts if barriers to adoption were better understood. This qualitative study assessed the needs of organizational stakeholders for a Caring Contacts informatics tool through interviews that investigated barriers to adoption, workflow challenges, and participant-suggested design opportunities. We identified contextual barriers related to

environment, intervention parameters, and technology use. Workflow challenges included time-consuming simple tasks, risk assessment and management, the cognitive demands of authoring follow-up messages, accessing and aggregating information across systems, and team communication. To address these needs, we propose design considerations that focus on automation, cognitive support, and data and workflow integration. Future work will incorporate these findings to design informatics tools supporting broader adoption of Caring Contacts.

3.1 Introduction

Suicide is a major public health concern [30–32]; at the same time, our lives increasingly unfold online, making new data sources available that enable the development of new supportive technologies. A range of suicide-related risk prediction models have been developed by the informatics and computer science communities [92,150], but little is known about how to integrate such models to support clinical practice at the point of care. This study endeavors to take these technologies a step closer toward translational impact by identifying opportunities for informatics support within an evidence-based suicide prevention intervention.

Caring Contacts is an effective suicide prevention; however, it has not been widely adopted. Organizations wishing to implement Caring Contacts face challenges in deciding how to make judicious use of available resources, address patient safety, and reach recipients in a meaningful way. Informatics approaches hold promise for addressing these challenges. The Caring Contacts intervention is potentially labor-intensive, such as when interventions with two-way communication yield responses indicating that the recipient is experiencing distress or an acute crisis. Programs must have an appropriate response and safety plan to reliably provide timely support in such cases. A promising area of investigation for informatics is, therefore, how to leverage suicide risk prediction models to triage responses to Caring Contacts messages.

However, as recommended by published frameworks that guide the operationalization of predictive models for clinical decision support (CDS), a thorough understanding of the context

of use and user needs must be established before model development begins [14,15]. Carefully designing informatics tools with the opinions and experiences of organizational stakeholders in mind is critical to successful adoption [151]. Therefore, there is a need to engage stakeholders to understand how an informatics tool can help address barriers to the adoption and delivery of Caring Contacts among healthcare organizations.

To address this research gap, we drew on principles of human-centered design [151] to describe specific barriers to Caring Contacts adoption faced by healthcare organizations, delineate workflow challenges faced by intervention staff, and formulate design considerations that can guide the development of informatics tools supporting the Caring Contacts intervention. The objective of this study was to characterize specific barriers to adoption, workflow challenges, and implementation bottlenecks among organizational stakeholders, including program coordinators, leadership, social workers, and intervention staff, affecting the Caring Contacts suicide prevention intervention. Findings inform design considerations for the development of informatics tools that help address these barriers.

3.2 Methods

We conducted a needs assessment using qualitative interviews to inform design considerations that meet user needs by directly engaging organizational stakeholders. The methods used for this study follow the principles of human-centered design as outlined by Maguire et al. [151] to establish the context of use and user needs. The Institutional Review Board of the University of Washington approved study procedures.

To understand how an informatics tool can help address the barriers to the adoption and delivery of Caring Contacts, we engaged organizational stakeholders with experience in planning, implementing, and delivering Caring Contacts. We purposively sampled interview participants with diverse perspectives as professionals in various roles, including program coordinators (e.g., principal investigators, care directors) and Caring Contacts authors (social

workers, psychologists, clerical staff). We recruited participants from programs using different message modalities (i.e., mail, email, text message) and in different intervention settings, including research (i.e., intervention research such as clinical trials) and clinical settings (i.e., routine primary care, specialty care, and public health programs). We recruited through the authors' existing professional networks of suicide prevention researchers and practitioners. We recruited professionals conducting suicide prevention programs serving marginalized groups, including Native American, rural, veteran, and active duty military communities. These recruitment efforts yielded mostly program coordinators, so we employed snowball sampling to reach additional participants in social work and clerical staff roles.

Table 3.1 Interview guide topics and example prompts

CC = Caring Contacts

Topics		Example prompts
A. Overall intervention structure, goals, and high-level challenges	Goals and expectations	What does your program seek to accomplish? How do you align patient expectations with the intervention goals? What related risks are there? How are they addressed?
	Barriers to intervention adoption	From your perspective, what's the biggest reason why CC is not more broadly adopted? How is the intervention funded? What resource limitations impact the intervention?
	Team makeup and dynamics	What roles have to be fulfilled to support CC? Who do you work with to support CC? How do you communicate with team members?
B. Task-specific challenges and corresponding design opportunities	Workflow challenges by task	Which tasks take the most time? Which tasks are most difficult? How do you solve the problems involved with these tasks? What do you need to complete the individual tasks (people, information)?
	Design opportunities by task	How could an informatics platform assist with these challenges? What are the most important things you require from a CC information system? Which tasks could be automated, and which tasks should not be automated?

Data was collected via semi-structured interviews. The interview guide was developed based upon Caring Contacts implementation challenges in prior research (Table 3.1). Across

three large randomized trials Comtois et al. [40,49,50] refined Caring Contacts via text message and, based on expert consensus, collected an initial list of challenges and bottlenecks that a digital solution might help address. Topic areas for inquiry were based on this expert input and organized around the structure of the workflow (i.e., eligibility determination and enrollment; scheduling and sending caring contacts; monitoring for incoming patient responses; determining urgency and how to follow up; authoring follow-up messages; creating any external documentation). Interviews were structured to flow from general intervention considerations to technology-specific challenges and design opportunities to identify where informatics tools might facilitate adoption and cost-effective, time-efficient delivery of the intervention. The interview guide was pilot tested with AK, who served as domain expert and interviewee, due to her experience as both a social worker serving as a Caring Contacts author and as a researcher acting as a champion.

Interviews lasted 45-60 minutes and were conducted via video conference by HAB, a Ph.D. student. The video conference software recorded and transcribed interviews for qualitative analysis. No bias due to a power differential was expected due to the relative seniority of participants and the absence of a professional relationship between interviewer and interviewee. We conducted interviews until reaching saturation, i.e., until no new themes emerged [152].

In accordance with guidance from Ancker et al. [153], we followed a four-stage process to conduct a deductive qualitative data analysis. First, we developed a codebook based on the topics for inquiry identified from expert input (KAC). Second, to enhance the reliability of coding, two authors (HAB, ML) independently coded one-third of transcripts. Codes applied by HAB and ML were compared in consensus meetings. The definitions for the codes in the codebook were iteratively refined based on feedback from the consensus process until there was agreement in coding between coders. Codes were applied to all interview transcripts by one author (HAB) [154,155]. Third, to discover themes within our topics of inquiry, coded excerpts

were grouped by similarity into emerging themes [155]. Fourth, we utilized member checking to verify the face validity of the themes.

3.3 Results

Sixteen individuals completed interviews (P1-P16), representing 12 unique Caring Contacts programs (Table 3.2). The organizations included large health systems (research and ongoing programs), community-based health advocacy groups (ongoing programs), military (research program), and a managed care organization (ongoing program). The populations served by the programs included both rural and urban communities, veterans, active duty military, and indigenous communities. Several programs sent caring messages via multiple modalities, depending on patient preference. Of the 14 participants (88%) who completed the demographic survey, nine (64%) were female, four were male (29%), and one was non-binary (7%). Ten identified as white (71%), four as Asian (29%), and one as Hispanic or Latino (7%). Six (43%) were mid-career professionals aged 40 to 49; six (43%) were 39 or younger, and two (14%) were older. Two (14%) worked primarily with indigenous communities and one (7%) was a suicide prevention professional with lived experience of suicidal thoughts and behaviors.

Table 3.2 Participant characteristics

		Participants (N=16) n (%)	**Programs (N=12) n (%)**
Role	**Coordinator**	6 (38%)	
	Author	5 (31%)	
	Both coordinator and author	5 (31%)	
Setting	**Research**	11 (69%)	7 (64%)
	Clinical (ongoing care)	5 (31%)	4 (36%)
Modality	**Mail**	6 (38%)	4 (36%)
	Text	11 (69%)	4 (36%)
	Email	5 (31%)	7 (64%)
	Phone	1 (6%)	1 (9%)

We report findings from interviews across the two topic areas in our interview guide: (A) *barriers and facilitators* in the overall work system surrounding the intervention structure,

goals, and high-level challenges involved with adoption, implementation, and overall success of current Caring Contacts programs, and (B) *challenges and their potential solutions* surrounding the day-to-day tasks of the workflow.

3.3.1 Overall work system barriers and facilitators

Three themes reflect workflow barriers expressed by participants: *Context and environment*, *Intervention parameters*, and *Technology*.

Context and environment. This theme describes issues stemming from Caring Contacts intervention settings, ranging from incentive structures, policies, and public health trends to business considerations at the healthcare organization level. Several participants mentioned difficulties in obtaining organizational buy-in. For example, P15, a program coordinator with lived experience of suicidal thoughts and behaviors, was concerned that compared with conventional approaches to mental health treatment, Caring Contacts could fuel skepticism:

> *"Initially, there was a lot of resistance … I was working on trying to find sustainability funding, and then it was like no we don't want to do this … I think part of that is the uhm, not really truly believing in the fact that peer support can make a difference, the way other support cannot. And so I think there's some of the older school thinking behind that decision making." (P15)*

Resource scarcity was a consistent theme limiting implementation efforts, ranging from a lack of funding for programs to insufficient staff, staff time, or necessary expertise. For example, P7, a Caring Contacts author on a research program, shared staffing and funding constraints:

> *"I think it would be great to offer [Caring Contacts] as another support system, but we also know that takes someone to do them, right. I mean it's going to be part of someone's*

workload, or maybe them being hired in for just that purpose, so it would really depend on whether or not they have the funding." (P8)

Other barriers at the organizational level include the lack of focus on prevention and mental health in current incentive structures, resulting in difficulties aligning unreimbursed prevention efforts with organizational priorities. For instance, P10, a message author working in clinical care, shared challenges with insurance reimbursement:

"How do we stop people from falling through the cracks in the medical field? They come in for something that's medical, and unfortunately society says that medical and mental health are two separate things and they're not. They are one big element that we should be treating the same, but instead insurance says, you can go to the hospital for your heart, but if you're having mental health issues we're not paying for therapy." (P10)

P9, a researcher, emphasized that the litigious nature of the current healthcare landscape requires carefully defining the scope of practice for liability reasons:

"Someone asked me a question like, I think it was a psychiatrist who asked me, … I'm providing my care, but once they're discharged my relationship with that person it's over like that legal clinical relationship is over. Are you asking me to like maintain this kind of legal you know clinic relationship past that, and I was like I kind of like - well, well yeah, so, not to, maybe. You know it's a valid question, and it does raise some issues of like liability too right. … Valid questions." (P9)

In addition to barriers to implementing Caring Contacts, participants also described several contextual and environmental facilitators. For example, organizational culture, such as an organizational mission to reduce suicides and attempts in the served population, resulted in healthcare organizations prioritizing suicide prevention. P10 shared:

"Our organization believes in it so much that they fund my position fully… we want to implement [Zero Suicide] fully, they have completely backed me … whatever I feel like I need to do for our patients to help them." (P10)

The expectation that suicide prevention will save money in the long run facilitated investments in Caring Contacts for organizations with payment structures that incentivize cost-effective care, as P13, a community health clinic director, shared:

"When you look at return on investment, we can also say that the savings do accrue to the health plan. Again this isn't why we did [Caring Contacts], but if you are a hospital, for instance, you could say, well, yes, this is good because it reduces suicidality, you know completed suicide, suicide attempts. But the savings don't actually accrue to the hospital, or to an outpatient provider. Ethically, morally, clinically it's the correct thing to do, but, in our case – again, this wasn't the driver – but we should see a financial benefit from reduced hospitalizations and ED visits" (P13)

Finally, P11, a coordinator and message author in a clinical care program, added regulatory requirements as another incentive for Caring Contacts:

"I think a big piece of what helped our system get to where it is, is the Joint Commission requirements as far as addressing Suicide Prevention and assessment" (P11)

Intervention parameters. Participants also described barriers and facilitators regarding intervention parameters, i.e. the intervention design and implementation specifics. This includes program goals and operating procedures, such as how people will be referred to the program, eligibility criteria, the number and timing of messages, the content of the initial Caring Contacts messages, whether to include disclaimers, and whether the program entails one-way or two-way communication. Considering the wide variety of intervention designs reported, each organization must carefully consider its individual approach, which takes time and effort. Establishing policies and communicating expectations that attenuate potential risks

of the intervention causing harm was a primary objective some participants described. For example, P3, a program coordinator and message author in a research study, described the challenges of developing efficient referral procedures:

"Another barrier …, in terms of kind of the complexity of moving information from the health system to the hotline, is just figuring out what that process looks like. So we have over 900 referring providers that are helping get patients referred to [the study], so training all of the providers that this resource exists and that it's available." (P3)

P5, a message author in a research study, added:

"I think having good protocols in order to respond well over text is probably the most time-consuming thing." (P5)

Establishing expectations with patients was mentioned as an essential component of avoiding patient harm. P7 shared the need to clarify with patients what types of support the program is or is not designed to deliver:

"We tell participants we're not a crisis service. And so we tell them we're not available 24 hours" (P7)

Further, adapting the intervention to the needs of specific populations and individual patients was time-consuming. Caring Contacts message authors must tailor the text, message schedules, and delivery modes for recipients of different age groups and cultural groups (e.g., indigenous communities, veterans, and healthcare workers). For example, P16, a program coordinator working with indigenous communities in a public health setting, shared:

"Our biggest lift is creating the messages [our population identifies with]" (P16)

P7 explained that messages should be caring and undemanding:

"Caring contacts is based upon the idea that if you feel connected to people in your community and you don't feel like a burden, you feel like you like belong, that your risk of

suicide goes down… It's never asking someone to do something or telling them to do something, it should just be like generally positive, encouraging good vibes." (P7)

P8 added that messages should not be too generic or repetitive:

"We don't like to repeat the same old thing, because it makes them feel like it's not personal for one thing, that like it's just computer-generated. And technically, it was software, but technically it was a human behind it." (P8)

The simplicity of Caring Contacts was seen as a barrier when the complexity of its logistics was overlooked. P4 shared:

"It's a combination of that it gets sold as a simple suicide prevention intervention. The concept is simple, right? It's mail. Yeah. The logistics of sending and managing a year's worth of mail is not simple." (P4)

Technology use. Technology was generally perceived as a facilitator, but was described as a barrier when it was difficult to use, introduced inefficient workflows, obscured needed information, or did not function correctly. For example, P5 reported that some software was not usable because it required specialized skills to operate:

"One of the reasons that we had a lot of issues with our Access Database is that it requires like special SQL code and stuff for it, that me and my coworkers are not specialized to program" (P5)

P12, a coordinator of a community-based program, described how software bugs necessitated labor-intensive workarounds:

"The capability to notify staff when a patient texts back … is really important. Because in the beginning, that email notification wasn't working. We ran into issues where I was seeing … no one responding back to this patient and it's been, you know, a day already. And I would reach out to the clinical care pointers and they're like Oh, we never got an email. … That actually happened, like, a few times. So that's why, like every day, I would monitor and also

have another clinical care coordinator, instead of relying on notification coming through … they would be in the portal, and they would check in the morning, midday, and then in the evening… those are kind of the type of bottlenecks that we did not anticipate" (P12)

Further, there were technology-related barriers arising from the characteristics and perceptions of the served population. Internet access and technology literacy were barriers to patient adoption of digital health tools, especially for elderly patients and those belonging to rural and native communities. For instance, P4 shared:

"There's a lot of disconnect … especially with age. Like I have folks… like, I have to walk them through the steps of how to use their phone. … I have one who doesn't… know how to send a text message, and he doesn't want to know how to send a text message either." (P4)

In contrast, technology was a facilitator when it eased data and workflow integration between different systems (e.g., transferring patient information between electronic health record (EHR) systems and text message platforms), provided mechanisms to plan and execute tasks (e.g., automatic message scheduling and sending), or captured data automatically (e.g., text message conversation history, call durations). For example, P3 shared how workflow integration with referring providers and data sharing across intervention staff facilitates the Caring Contacts intervention:

"We have a best practice advisory alert the fires in Epic which informs providers that you have a patient that may be eligible… We also built out through EpicCare Link the ability to share portions of patient charts with the hotline directly, so they can see directly the safety plan that was developed… there's a lot of technology that is facilitating this work for us." (P3)

3.3.2 Task-specific workflow challenges and participant-suggested design opportunities

Five themes reflect workflow challenges and design opportunities reported by participants: Time-consuming simple tasks, Managing risk, Authoring helpful follow-up messages, Accessing data across sources/systems, and Team communication and collaboration.

Time-consuming simple tasks. The workload imposed by suicide prevention efforts was a recurring theme, and is particularly concerning in the context of the resource constraints we identified as system-wide barriers. The burden due to inefficient workflows emerged as a significant bottleneck in both clinical and clerical tasks completed by Caring Contacts authors. Program coordinators are affected by this as well, because productivity levels will determine staffing requirements. Participants described considerable time spent on repetitive manual tasks, such as eligibility determination, scheduling and sending templated messages, and documenting communications. For example, P1 explained:

> *"Every person I enroll, I want them to get 11 texts on this schedule, you know, one day, one week, one month, two months. You couldn't just tell it that; you had to literally go in there and every single message, pick the date and time." (P1)*

On the other hand, P14, a Caring Contacts message author in a research study, shared that automation of such tasks is a big help:

> *"The fact that it just sends out texts for us automatically on a schedule and we don't have to manually type or send a text is obviously very helpful." (P14)*

P10 described the manual nature of documenting contacts and their hopes for future technological assistance:

> *"In our EHR, I write that a caring contact did go out and what date it went out. … I type it out because I'm the only one. I'm sure that, as the system hires more zero suicide*

coordinators, because that's the hope, that they may put something in that you can just hit a button for. But for now, I just type it out." (P10)

Managing risk. Participants reiterated that first and foremost, patient harm must be avoided. This requires vigilance regarding the content of outgoing messages, monitoring and assessing risk in patient responses, and composing follow-up communications. Managing risk is a shared responsibility between all stakeholders, and therefore, both Caring Contacts authors and program leadership/coordinators are affected. One key component is timeliness. Participants perceived intense pressure to provide high-quality support promptly, especially in urgent crisis situations. P3 shared the need to keep close track of patients:

"Just having a better way to track and make sure nobody falls through the cracks. I think all of us have a lot of anxiety about potentially causing harm by introducing this relationship and then someone drops accidentally when they really need help, I mean that's like the worst thing that could happen… it's a big concern." (P3)

P6 shared frustration about manually collecting relevant information when time is of the essence:

"Being on call, you know… if I get a text message at, you know, one o'clock in the morning, and it's something distressful, and I don't even know who this person is because it wasn't one of my [patients], now I've got to, you know, … gather my wits about me, trying to get on it, get some insight and some direction, and to be able to respond as quickly as possible and as accordingly as possible to ensure that person gets connected and is safe." (P6)

Reviewing historical interactions can also reveal sensitive issues. P1 shared that bringing up something that is difficult for the patient to think or talk about should be avoided:

"If you mentioned their access to clinical care, and like the last few times they wrote back … with you know some degree of distress you might not want to [mention their access to care] again." (P1)

P12 shared an approach to automatic flagging of potentially high-risk messages, which could help address risk management concerns:

"Say if a patient texts back and uses any of the words like, you know, hurt, kill, like any of those words, the system flags it … then sends an automated response text back to the member letting them know resources if they need help, and then also we get notified too … [we were] able to review that list and anytime we want to add new words to that list we can do that" (P12)

Authoring helpful follow-up messages. Apart from the scheduled, pre-written (or templated) Caring Contacts messages, the ongoing execution of the program may require being responsive to patients who reach out for further support. While most responses from patients are straightforward expressions of gratitude, there are rare patient responses that express distress and require careful consideration to determine follow-up actions based on each patient's individual situation. In these cases, participants emphasized that writing follow-up messages is a complex cognitive task. P1 shared:

"Authoring responses is a lot of work." (P1)

If the intervention team determines that a Caring Contacts recipient is experiencing adversity, the team must extrapolate the patient's needs (e.g., crisis support vs. encouraging words) and compose an appropriate follow-up message. Message authors may review the patient's history and known coping mechanisms. P14 shared:

"We look in their medical record and see if there is a safety plan … their supports and coping mechanisms and things like that." (P14)

Assessing patients and their communications accurately is another concern. P1 shared the potential need to incorporate patient-specific factors, such as a patient's baseline risk level, to avoid missing warning signs:

"There was one person that we classified as urgent where it kind of took a little nuance to get it, you know, for him it was urgent." (P1)

Finally, P16 added that a message author might customize message resources based on the patient's individual needs:

"If somebody responds back and is like, oh sorry, I got a really bad grade on my math test, so that's why I told you I was feeling down. And then we'll respond back and be like, oh that's such a bummer, did you know that, you know, if you ever want some more resources about math tutoring or anything like that to try out … From where you're located, here are some resources for that area, specifically." (P16)

Accessing data across sources/systems. Participants reported referring to information from several systems throughout the intervention workflow. For example, when evaluating a patient response or composing follow-up messages, staff might review prior message exchanges within the messaging system, but they might also refer to external information such as the patient's demographics, details of their clinical history (e.g., suicidality questionnaire responses, previously documented safety plans), their current healthcare providers, and past or upcoming appointments. These data may be in one or more EHR systems, intervention-specific records (e.g., patient notes kept by intervention staff), or other data sources (e.g., text-messaging platform). Caring Contacts authors are affected by these challenges in terms of productivity and ease of use. Program coordinators have to take these challenges into account when making intervention design and implementation decisions, balancing feasibility and affordability with ease of use and intervention reach. For example, P11 shared that EHR data can be informative:

"Risk factors are broad that could be someone having particular diagnoses …. [or] they're missing a bunch of appointments" (P11)

P6 shared that integrating patient data in a single access point is desirable:

"If there was some way to provide all of that in a centralized location, where maybe … you could write notes or keep a log of all the service members versus having to go to an excel spreadsheet and open up, and like kind of toggle back and forth, if you could have that kind of functionality … with one application" (P6)

P11 shared that access to and integration of different systems is a benefit:

"The benefit of us having a work system is that a lot of who we're working with is kind of integrated into our current system, so I can typically see the documentation of the therapist or the treatment team, or whoever it may be. And also communicate with them in the electronic health record. Our behavioral health system uses one and our like medical side, like the hospital and stuff they use a separate one, but we can access them both." (P11)

Team communication and collaboration. Caring Contacts team members have different responsibilities, expertise, and roles, ranging from behavioral health providers and social workers to program administrators and clerical support staff. Participants describe collaboration across team members as critical to completing intervention tasks requiring diverse expertise. This affects all Caring Contacts authors. For example, P16 described routing messages to follow-up specialists based on content:

"If it's a concerning thing, he will route it to our concerning message team, within our protocol for what to do whenever we receive a concerning message/post/email. And depending on why it's concerning, it goes to certain people." (P16)

Sharing information across the team is essential for efficient handoff. P10 shared that different staff collect patient information at different points in a patient's journey, requiring the transfer of notes:

"Our behavioral health evaluators are the ones that will meet them in the hospital and do all of the treatment and engagement with them in the hospital. … they will print off a face

sheet and send that to me, and sometimes they will write notes like, dealing with this, please send them extra cards, or call and send the card" (P10)

Similarly, sharing intervention notes and updates with referring providers can benefit patient care continuity. P14 shared challenges to data sharing between providers due to poor data integration and system access:

"The notes that we write are kept in the research database … and their providers don't have access to that. If they create a safety plan with their provider or with a social worker … then that is included in their chart and is visible to [both us and] their provider, but if we create the safety plan, their provider doesn't have access to that." (P14)

Finally, sharing the workload between intervention staff is necessary, especially when staff cover for each other (e.g., after hours). Thus, team communication is essential and represents a barrier when inefficient or when misunderstandings occur. P11 shared the need to communicate how the workload is balanced across team members:

"Just some sort of way to know that someone else did it even if it's not fully documented yet…. That's something that we run into as a team, that we try to support each other, but if we don't get the opportunity to communicate something like that, it just, it doesn't go as smoothly." (P11)

3.4 Discussion

Through this needs assessment, we established the context of use, identified barriers to adopting the Caring Contacts suicide prevention intervention among organization stakeholders, and identified opportunities for informatics tools to help address these barriers. Our findings broadly fall into two categories: overall work system barriers and facilitators and task-specific workflow challenges.

Any informatics tool must be designed with the context of use in mind. The work system themes that establish this context included high-level contextual and environmental obstacles,

intervention design and implementation issues, and system-level technology concerns. While it may not be possible to address all barriers and challenges directly with an informatics tool, they must be accounted for, as they place constraints on tool design and deployment. For example, a significant limitation is that the incentive structures currently in place in the U.S. healthcare system are not favorable to preventive care. To make Caring Contacts broadly feasible within this constraint, we must prioritize the judicious use of human resources to control costs.

Task-level workflow challenges represent pain points in current workflows and therefore reflect opportunities for informatics tools to help. Five themes emerged from these workflow challenges. We now present design considerations for addressing these challenges within the work system constraints, including context of use/environment, intervention design, and technology use, along with examples illustrating how informatics support could be implemented in future work.

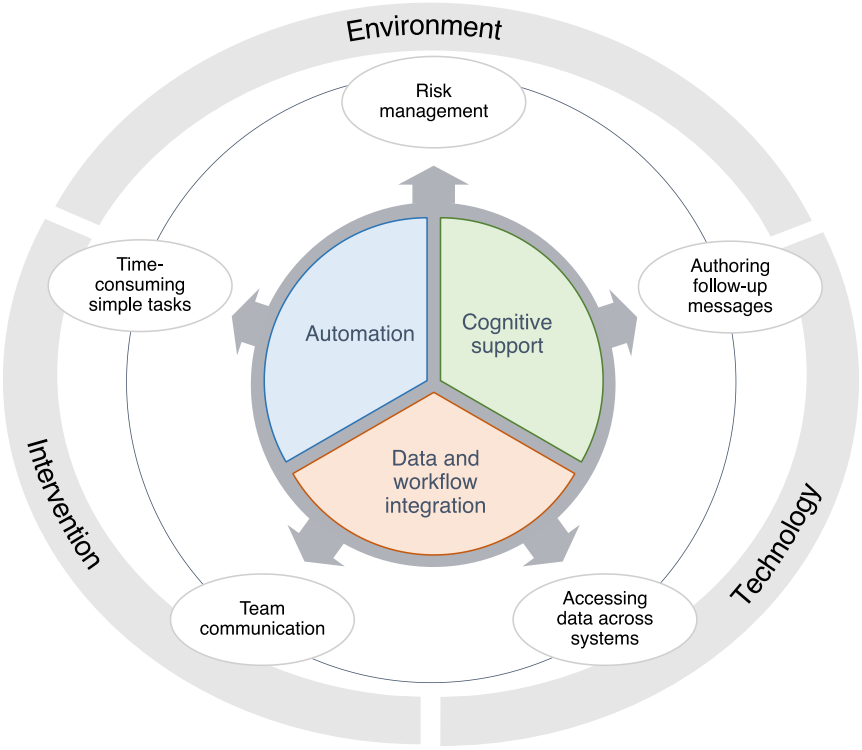

Figure 3.1 Summary of findings mapped to design considerations for informatics-supported suicide prevention. Work system constraints are shown on the outer circle. Workflow challenges are shown in bubbles. Design considerations for addressing challenges are shown in the inner circle.

Our findings point to three cross-cutting design considerations that are key to addressing these workflow challenges: *Automation, Cognitive support*, and *Data and workflow integration* (Figure 3.1).

Automation. Resource scarcity emerged as a recurring theme. The workload imposed by Caring Contacts can present an insurmountable barrier to implementation. Informatics tools can help alleviate this burden, as many repetitive tasks could be partially or fully automated. These tasks include checking patient suitability or eligibility for Caring Contacts, scheduling initial messages, and sending automated responses to texts or emails. There are also opportunities to integrate predictive modeling tools. For example, several promising approaches to personalized risk modeling using Natural Language Processing (NLP) and machine learning have been reported [92,150]; such methods could be used to trigger automatic responses with a crisis line phone number to a patient who expresses immediate intent for self-harm and to flag and prioritize messages for follow-up. Messages could also be automatically routed to staff with appropriate expertise based on message content.

Information retrieval and synthesis for cognitive support. Authoring follow-up messages can be a demanding cognitive task. For this purpose, intervention staff must not only collect various pieces of relevant information, but also synthesize and harmonize that information in order to make clinical judgments and formulate follow-up actions. Participants described using different kinds of data and insights, including a recipient's health history (e.g., demographics, diagnoses, appointments), risk and protective factors (e.g., social determinants of health), and interaction history (e.g., to determine what kinds of messages were previously well received). Informatics tools could provide cognitive support by aggregating and synthesizing information from different data sources in an easily digestible format (e.g., a timeline of challenging life events and corresponding trends in suicide risk) to reveal insights for CDS. A system could also draw links between available resources and patients' resource needs and present suggestions to authors accordingly. Opportunities for advanced cognitive support

tools include mining message exchange histories for patterns in patient communication and automatic extraction of clinically relevant insights and therapeutic opportunities (e.g., symptoms, risk factors emerging from patient communications, warning signs).

Data and workflow integration. The intervention workflows that informatics tools for Caring Contacts should support require communication between systems and people with different roles and responsibilities. For example, programs may aim to provide some level of support around the clock, which means that follow-up specialists may not know individual patients well (e.g., when covering for someone else); Caring Contacts tools may draw on previously described technology-facilitated handoff approaches [156] to support these workflows. Additionally, it is essential to avoid miscommunications, e.g. to prevent harm due to a patient in need of support not receiving a response due to a lack of clarity regarding patient assignments. Tools should therefore provide planning tools such as task and priority lists. Information exchange with other health information technology platforms is also critical. Automatically incorporating external patient information, e.g. safety plans from the referring providers' EHRs, can benefit intervention staff as they evaluate and follow up with patients, without requiring duplicated data entry efforts. Similarly, sharing intervention notes and updates back to the referring provider's EHR could support ongoing care or insurance reimbursement. Data integration can also facilitate reporting on outcome metrics to help Caring Contacts program coordinators monitor and improve the program. Additionally, workflow integration is an essential component of reducing the workload imposed by working in multiple information systems. While data and workflow integration between health information technology platforms are well established requirements, they have been prohibitively difficult to enable for many reasons, including privacy and governance concerns, technical infrastructure, and diverging data formats [7,157,158]. Fortunately, health data standards and exchange protocols are maturing, and today we have well-adopted, freely available standards such as Fast Healthcare Interoperability Resources (FHIR) [26] and workflow integration tools such as

SMART-on-FHIR [25] and CDS Hooks [159] that make it possible to overcome some of these barriers. For example, SMART-on-FHIR could enable users to launch Caring Contacts workflows directly from within the EHR, automatically pulling in relevant patient record information. Additionally, FHIR includes API endpoints for common data exchange tasks; for example, with appropriate permissions in place, an external system can use a FHIR endpoint to automatically file a clinical note to any EHR implementing FHIR.

Recently, there has been a proliferation of suicide-related risk prediction models [150] using diverse data sources, e.g. patient-generated natural language data [92] from Facebook [95], Twitter [160], and Reddit [161] posts, query terms used for internet searches [94], and EHR data such as diagnoses [115]. These approaches demonstrate promise, but it is unclear how to operationalize such models to make a clinical impact, which motivates the current work. CDS tools have been intensely investigated in the field of biomedical informatics, and guidelines for designing effective decision support have been published [162]. Based on these guidelines, critical questions for CDS tools for Caring Contacts include: Who has the ability to act upon risk predictions in a way that impacts outcomes? When and how can they benefit from predictions? At that point, what data is available for inference? What else is needed for a clinician to act upon the new knowledge within their established workflows? These questions illustrate the complexity of creating decision support tools with the potential to improve clinical care and impact patient outcomes, which is further underscored by the scarcity of examples of success described in the literature. To guide such work, Shah et al. [15] developed a framework for making predictive models useful in practice. Here, we aimed to complete the indispensable first step described in this framework: establishing the use case for a suicide risk assessment model. Our work identified several use cases for artificial intelligence tools, and helps answer the questions listed above. It also revealed a broad set of challenges and opportunities related to end-to-end workflow support, suggesting the need to adopt a broader perspective. In order to realize the translational impact of new and existing predictive technologies, we must therefore

design a CDS system that supports the workflow comprehensively. Our design considerations provide holistic guidance for both the workflow and artificial intelligence components that Caring Contacts informatics support tools may include. In future work, we will use these findings to design and develop an informatics tool suitable for a pilot deployment.

This work must be considered in light of its limitations. Our participants were sampled from our existing professional network and subscribers of a suicide prevention mailing list who responded to a volunteer request, which may have biased our sample toward those who already perceive more benefits than barriers to Caring Contacts implementation. While we tried to capture broad perspectives on barriers, our focus on technology may have biased participants' responses. Finally, this work investigated the perspectives of the potential users of Caring Contacts information technology who administer the intervention, rather than the perspectives of patient users who are the targets of the intervention. Prior work has investigated the acceptability of the intervention to patients [40,43]; however, if the use of novel informatics tools were to change the patient experience of the intervention, it would be necessary to re-engage patients to ensure acceptability.

3.5 Conclusion

This work identified barriers to adoption, workflow challenges, and design opportunities for informatics tools supporting the Caring Contacts suicide prevention intervention among organizational stakeholders. With newfound clarity regarding the opportunities for technology support, including CDS tools such as risk prediction models, this work contributes to realizing the translational potential of informatics interventions to benefit clinical care.

Chapter 4. Behavioral activation and depression symptomatology: Longitudinal assessment of linguistic indicators in text-based therapy sessions

Aim 1 of this dissertation revealed that follow-up message authoring can be a cognitive burden in the Caring Contacts intervention, in part due to the need to review patients' message histories for content that should be taken into account. There is therefore an opportunity for an AI-based tool to automatically extract such content items to lighten this burden. For the purpose of clinical decision support, it is imperative that extracted insights be actionable. Clinically validated theoretical constructs can serve this purpose, as long as they not only inform the clinical understanding of a patient's pathology, but also intervention strategies.

In the work described in this chapter, my co-authors and I investigated the use of LIWC and distributional semantics to automatically extract clinically relevant insights in a large corpus of patient-generated text, focusing on behavioral activation therapy for depression. Reduced behavioral activation, i.e. participation in rewarding activities, is a hallmark of depression; therapy may aim to induce such behaviors, triggering a positive feedback loop that leads to symptom improvement. Reduced behavioral activation is therefore both a symptom and an intervention target. Currently, therapists utilizing behavioral activation therapy measure symptoms and therapy progress by regularly administering a validated questionnaire instrument, which interrogates the seven dimensions of behavioral activation: satisfaction with activities, breadth of activities, autonomous decision-making regarding activities, deriving a sense of accomplishment from achieving activity-related goals, planning for long-term goals, enjoyment of effort, and structuring daily activities. These seven items are therefore clinically validated, actionable constructs. In this work, I developed and evaluated an approach to

automatically measuring these dimensions of behavioral activation using patient-generated natural language.

Therefore, this work is uniquely relevant to informatics-supported suicide prevention. The constructs of behavioral activation are relevant to depression, and may be directly applicable because depression is closely related to suicidal thoughts and behaviors. However, in future work, my approach could also be used to develop ways to capture clinically actionable constructs specific to suicide, such as hopelessness or entrapment.

Burkhardt HA, Alexopoulos GS, Pullmann MD, Hull TD, Areán PA, Cohen T. Behavioral Activation and Depression Symptomatology: Longitudinal Assessment of Linguistic Indicators in Text-Based Therapy Sessions. *J Med Internet Res*. 2021;23(7):e28244. doi:10.2196/28244

Abstract. Background: Behavioral Activation (BA) is rooted in the behavioral theory of depression, which states that increased exposure to meaningful, rewarding activities is a critical factor in the treatment of depression. Assessing constructs relevant to BA currently requires the administration of standardized instruments, such as the Behavioral Activation for Depression Scale (BADS), which places a burden on patients and providers, amongst other potential limitations. Previous work has shown that depressed and non-depressed individuals may use language differently and that automated tools can detect these differences. The increasing use of online chat-based mental health counseling presents an unparalleled resource for automated longitudinal linguistic analysis of patients with depression, with the potential to illuminate the role of reward exposure in recovery.

Objective: This work investigates how linguistic indicators of planning and participation in enjoyable activities identified in online, text-based counseling sessions relate to depression symptomatology over time.

Methods: Using distributional semantics methods applied to a large corpus of text-based online therapy sessions, we devised a set of novel BA-related categories for the Linguistic Inquiry and Word Count (LIWC) software package. We then analyzed the language used by 10,000 patients in online therapy chat logs for indicators of activation and other depression-related markers using LIWC.

Results: Despite their conceptual and operational differences, both previously established LIWC markers of depression and our novel linguistic indicators of activation are strongly associated with depression scores (PHQ-9) and longitudinal patient trajectories. Emotional tone, pronoun rates, words related to sadness, health, and biology, and BA-related LIWC categories appear to be complementary, explaining more of the variance in the PHQ score together than they do independently.

Conclusions: This study enables further work in automated diagnosis and assessment of depression, the refinement of BA psychotherapeutic strategies, and the development of predictive models for decision support.

4.1 Introduction

Over 20% of adults in the United States have a mental illness [32]. Depression is among the most common mental health disorders: over 19 million adults suffered major depressive episodes in 2019. Effective delivery of mental health services is a challenge for many reasons, including that individuals respond differently to therapy [163,164]. To maximize treatment benefits, mental health care providers must continually assess progress and adjust treatment plans [163]. From a research perspective, longitudinal information about known and hypothesized mechanisms of recovery is a prerequisite to the refinement of current

interventions and can inform the development of new ones. Validated survey instruments exist to assess symptoms and other constructs relevant to therapy delivery and progress [165]; however, repeatedly filling out questionnaires places a burden on patients and providers, limiting the frequency with which these data can be collected. In contrast, using already available data created as part of routine care obviates the need for additional data collection. Additionally, it has been argued that subjective self-reports present potential limitations, for example due to cognitive and memory bias [166]; while careful scale design can alleviate these problems, objective, naturalistic measurements may be preferable.

4.1.1 Behavioral activation and engagement

The behavioral theory of depression states that depressed individuals participate in fewer pleasant activities and derive less pleasure and feelings of accomplishment from such activities [167]; in other words, they exhibit reduced behavioral activation. This phenomenon is self-exacerbating: reduced activation represents a loss of positive feelings that further reduces activation. Neurobiological findings suggest that dysfunction of reward networks (especially in Reward Valuation, Effort Valuation, Action Selection, Preference-based Decision-making, and Reward Learning) is a central process perpetuating depression [168,169]. For this reason, "reward exposure" aiming to induce behavioral activation has been thought to reactivate and retrain reward networks and improve depression [170]. Behavioral Activation (BA) therapies are therapeutic approaches based on the relationship between depressive symptomatology and engagement with pleasant activities. They aim to reduce depression symptoms by activating the reward system, and have been shown to be as effective as learning-based therapies while being easier to understand for patients and easier to deliver for therapists [167]. An example is the streamlined, evidence-based psychotherapeutic strategy called Engage [168], which aims to systematically address disengagement from participation in pleasurable activities in a structured approach by incorporating reward exposure and addressing barriers in three behavioral

domains: negativity bias, apathy, and emotional dysregulation. A recent randomized controlled trial showed that Engage is as effective as problem-solving therapy in treating late-life major depression, while having the advantage of being less complex; Engage required 30% less training time compared to problem-solving therapy [171].

To better understand the relationship between BA-based therapies and therapeutic response in depression, robust metrics of the underlying theoretical constructs that minimize reliance on patients' and providers' subjective reports are needed. Text-based therapy sessions provide a unique opportunity to develop such metrics because all language exchanged in these encounters is archived.

4.1.2 Language as an indicator of mental state

Language reflects both conscious and subconscious thoughts and feelings [89–91]. Previous work has shown that depressed individuals use language differently than non-depressed individuals in a manner anticipated by cognitive theories of depression. For example, depressed individuals use more first-person singular pronouns ("I", "me", "my", etc.) than non-depressed individuals [172,173], indicating increased self-focused attention, a language use consistent with Pyszcynski and Greenberg's integrative model of depression [174]. Depression has also been shown to be associated with a lack of social integration or social disengagement [175–177]. For this reason, Rude et al. [172] anticipated a reduction in use of first-person plural pronouns ("we", "us", etc.) in depressed college students, but had too low a base rate to assess its impact in the sample available for analysis. Stirman et al. [177] found that suicidal poets used fewer first-person plural pronouns than nonsuicidal poets. Linguistic indicators of positive and negative affect differ in depression, and have shown utility in social media-based predictive models of depression [123,172]. These findings are consistent with the emphasis on negative valence in Beck's influential depression theory [178]. Finally, prior work has investigated content word usage by depressed individuals compared to control groups without depression.

These include words related to sadness, as well as words related to somatic health concerns (health and biology words, with the biology category combining body, health, sexual, and ingestion words) [129,133]. Given these findings, the question arises whether variations in language use related to the behavioral theory of depression can be detected through natural language processing.

One approach to capturing emotional affect, linguistic style, and topics in written text is to calculate the percentage of words belonging to defined categories, such as positive affect words, pronouns, or words related to certain topics, e.g. health or leisure. The Linguistic Inquiry and Word Count (LIWC) software package, a tool developed to study linguistic indicators of mental states, embodies this technique and was used to quantify relevant pronouns and affect words in the work discussed above. As reviewed by Pennebaker et al. [122], numerous experiments have validated LIWC's categories. However, while LIWC constructs such as 'leisure' are related to the notion of activation, they do not provide a comprehensive account of how engagement might manifest in language. For example, categories of relevance to BA, such as the breadth of activities one engages in or the extent to which one derives a sense of accomplishment from setting and achieving activity goals, are not represented in LIWC's standard dictionaries.

Distributional representations of words learned from large amounts of electronic text can help construct comprehensive sets of terms similar to the curated sets used by LIWC to represent categories. Also known as semantic vectors or *word embeddings*, these representations are learned from text, with a typical approach involving first initializing random vectors of user-defined dimensionality, and then iteratively updating them to make vectors for words appearing in similar contexts similar to one another. With neural embeddings, this is achieved by training a neural network model to predict the words surrounding an observed word, and retaining some of the neural network weights after training to serve as word embeddings. Empath [179] is a tool designed to support rapid computer-assisted construction of

user-defined term sets using such embeddings to find terms that are similar to an initial set of seed terms. Term sets constructed in this way have a strong correlation with the corresponding LIWC categories, which were constructed in a completely manual process. In essence, Empath's approach uses distributional representations of words to identify similar terms to a set of seed terms based on their distributional statistics across a large text corpus. In this way, a small seed set of terms can be rapidly expanded to provide adequate coverage, with the expanded list provided to manual reviewers for pruning of those terms considered to be inconsistent with the category of interest. Empath's vector representations are derived from a corpus of fiction. Though generally harder to come by, customized in-domain training corpora are known to produce better word representations in clinical domains [180].

For the current work, we developed a metric of behavioral activation, using distributional representations derived from a large corpus of naturally occurring language from online therapy chat messages (n=2,527,783), and characterized its relationship to indicators of depression severity. We hypothesized that linguistic markers of activation would be more frequent in milder depression than in severe depression and that longitudinal changes in these markers would reflect the trajectories of patients' depression; patients who improve over time should also show an increase in behavioral activation. We further hypothesized that linguistic markers of behavioral activation would capture a separate, clinically meaningful dimension of depression symptomatology – namely, engagement in meaningful, rewarding activities – compared to the established linguistic indicators, which capture psychological manifestations of depression (self-focused attention/social integration (function word usage) and emotional tone) and content topics (sadness, health, biology words). Therefore, the behavioral activation metric should capture information beyond that reflected by established markers. We tested these hypotheses in the subset of messages from the time period where evaluations of the severity of depression were available for participants at regular intervals (n=1,051,025).

4.2 Methods

4.2.1 BA lexicon

We developed a lexicon of related words, collectively representing the construct of activation as used in BA. We constructed a set of 66 unique representative seed terms, informed by the Activation subscale of the Behavioral Activation for Depression Scale (BADS) [181], a validated instrument used to identify subjective engagement levels. The subscale consists of seven questions, each aiming to capture a unique component of the construct. Seed terms were selected manually for each question in collaboration with G.A., a clinician investigator with extensive experience in BA approaches (Table 4.1).

Table 4.1 Seed terms derived by the authors from the individual questions on the "Activation" subscale of the Behavioral Activation for Depression Scale (BADS).

The name we assigned for each item (for brevity) is shown in parentheses. Note that there are 104 total words in the right column, including duplicates (e.g. "goals" appears in accomplishment, long-term, and structure), for a total of 66 unique terms.

Item	Derived seed terms
I am content with the amount and types of things I did. (satisfaction)	accomplish, achieve, satisfaction, satisfied, enjoy, content, contentment, accomplishment, love, proud, inspired, inspiring, enthuse, affirm
I engaged in a wide and diverse array of activities. (breadth)	activity, active, participate, involved, event, powerlifting, watercoloring, exercise, sport, basketball, restaurant, hobby, craft, art, music, instrument, piano
I made good decisions about what type of activities and/or situations I put myself in. (decisions)	decision, planning, plan, contest, competition, opportunity, chance, spontaneous, whim, spur, attentive, affirm, commit, focus
I was an active person and accomplished the goals I set out to do. (accomplishment)	goals, accomplish, progress, goal, achieve, effort, content, contentment, accomplishment, proud
I did things even though they were hard because they fit in with my long-term goals for myself. (long-term)	goals, progress, goal, effort, planning, plan, challenge, attentive, birth, commit, change, invest, life, payoff, benefit
I did something that was hard to do but it was worth it. (effort)	effort, enjoy, excited, energized, energizing, love, contest, competition, challenge, chance, fun, enthusiastic, inspired, inspiring, enthuse, event, affirm, commit, change, focus, fuel, invest, invigorate
I structured my day's activities. (structure)	goals, progress, goal, planning, plan, structure, attentive, event, routine, schedule, regular

We expanded the sets of terms for the novel LIWC construct by using methods of distributional semantics, which generate vector representations of words from their distributional statistics in text, such that words occurring in similar contexts will have similar vector representations [182,183]. Specifically, we used the open source Semantic Vectors software package [184–186] to train 100-dimensional word embeddings using the skipgram-with-negative-sampling algorithm [187] on a set of 2.5 million de-identified messages sent by Talkspace clients (>165 million total words). Embeddings were trained over 10 epochs, using a sliding window radius of 2 and a subsampling frequency threshold of 10^{-5}. Words occurring fewer than five times in the corpus were excluded from training. This minimum frequency threshold is employed to restrict model consideration to those terms that occur in a sufficient number of contexts to inform a distributional representation, and to constrain the number of vectors to save time (during nearest neighbor search) and disk space. We did not attempt to optimize this parameter, but note that it is the default in the canonical implementation of the skipgram-with-negative-sampling algorithm [188]. For each seed term, we then added the 30 most related terms as determined by the cosine similarity between the seed term's vector representation and the vectors for all other terms. We chose to add 30 because this number appeared to achieve high coverage while imposing a manageable workload for manual pruning. Note that stemming is not necessary, as this process will capture all forms of a word appearing in the training text, while preserving their semantic nuances; further, keeping all words appearing in the raw text ensures consistency between our dictionary and the texts to be assessed. For illustrative examples of similar words, see Table 4.2.

Table 4.2 Examples of seed terms and similar terms with corresponding similarity score, calculated by computing the similarity between word vectors.

[a]Terms were extracted from our chat message corpus and thus include common typographical errors. [b]Words that were removed in the filtering process.

Seed term	Similar terms[a]	Similarity score
Proud	Accomplished	0.729
	Accomplishment	0.679
	Accomplishments	0.673
	Impressed	0.667
	Prouder	0.663
	Gussied	0.646
Active	Inactive[b]	0.659
	Activity	0.633
	Powerlifter	0.607
	Motivated	0.605
	Mighy	0.600
	Intramural	0.592
Decision	Decisions	0.863
	Choice	0.841
	Deciding	0.723
	Hyphenating	0.697
	Choices	0.693
	Decide	0.671
Goal	Goals	0.865
	Attainable	0.761
	Achievable	0.740
	Acheive	0.725
	Aim	0.722
	Accomplish	0.717
Commit	Committing	0.828
	Committed	0.788
	Babydaddy[b]	0.708
	Committ	0.708
	Commitment	0.706
	Sucide[b]	0.682
Effort	Efforts	0.729
	Concerted[b]	0.718
	Valiant	0.690
	Handsomeness[b]	0.687
	Timeand[a]	0.662
	Independents[a]	0.648
Routine	Routines	0.874
	Schedule	0.708
	Nighttime	0.698
	Regimen	0.691
	Rhythm	0.682
	Schefule	0.682

These lists of terms were manually filtered to remove irrelevant or inaccurate terms. Then, we solicited feedback and suggestions from G.A. Feedback was incorporated at the seed term and manual filtering steps. The process was iteratively repeated until the lexicon was found to be satisfactorily inclusive and specific. Finally, the expanded lists of words, one per seed term, were combined to form seven partially overlapping sub-concepts according to Table 4.1, as well as one overarching *activation* category. In the overarching activation category, duplicate terms were removed to prevent double counting. We obtained a set of 1059 unique words, which represent the overarching idea of behavioral activation. The words originating from each of the seven items in the activation subscale of the BADS yielded the sub-concepts: *satisfaction* (227 words), *breadth* (341 words), *decisions* (205 words), *accomplishment* (154 words), *long-term* planning (240 words), enjoyment of *effort* (342 words), and *structure* (216 words). LIWC was then used to measure the frequency of words belonging to each construct as a metric of patient engagement in behavioral activation.

4.2.2 LIWC

The Linguistic Inquiry and Word Count (LIWC) [121,122] software package, developed by Pennebaker and his colleagues over the past two decades, was used for linguistic analysis. LIWC derives features from narrative text by counting the number of words in a text that correspond to categories in LIWC's lexicon (or dictionary), with categories defined by lists of words that fall into them. LIWC returns the percentage (or proportion) of words in a text that correspond to each category. For example, consider the following excerpt from an interview with singer, songwriter and poet Leonard Cohen:

> *"When I speak of depression, I speak of a clinical depression that is the background of your entire life, a background of anguish and anxiety, a sense that nothing goes well, that pleasure is unavailable and all your strategies collapse."* [189]

This excerpt is 40 words long, and the words' depression' (n=2), 'anguish' (n=1) and 'anxiety' (n=1) fall into LIWC's negative emotion category. Therefore, LIWC returns a percentage score of 10% (100*4/40) for this category. Other categories are measured similarly by estimating the frequency with which words they include occur in a unit of text. However, LIWC also includes a set of composite categories that are derived by combining individual categories. As negative and positive affect are both potentially informative, for parsimony, we considered the composite emotional tone variable, which combines the positive and negative emotion categories. A high tone score indicates a predominance of positive over negative emotion words, and a low score indicates the opposite. A score of 50 indicates a balance between positive and negative affect [121].

Additionally, we measured the usage rates of first-person singular pronouns, first-person plural pronouns, and words belonging to the content categories health, biology, and sadness. Finally, we measured linguistic indicators of behavioral activation by counting the number of BA lexicon words overall, and in each subcategory, in every patient's messages.

4.2.3 Data

This work utilized de-identified chat messages sent during routine online therapy, collected for a previously reported study by Hull et al. [190]. Clients took part in messaging therapy, conducted by a licensed, certified clinical professional via the Talkspace online platform, over 12 weeks. The platform provides a paid service open to all, and the service may be covered by some insurers. Therapists and clients converse via written, asynchronous messaging on the platform, and therapists utilized a range of therapeutic strategies. The platform also allows users to send video and audio messages, though these were not used in the current work. Only client messages, collected during the course of therapy, were used in this study. Participants completed Patient Health Questionnaire 9-item (PHQ-9) questionnaires at baseline as well as every three weeks during therapy. The PHQ-9 is a validated self-report questionnaire

commonly used to assess depression severity, scored on a scale of 0-27 [165]. For further details on the platform, data collection process, and study population, see Hull et al. [190].

The participants (N=10,718) were young (79% 35 years old or younger; none under 18), educated (74.9% Bachelor's degree or higher), and mostly female (78.9%). Data on race and ethnicity are not systematically collected by the digital platform and are missing for most participants. There were a total of 24,387 PHQ-9 assessments with corresponding messages, with 37.6% of participants (n=4,035) only completing the baseline assessment, 24.5% (n=2,626) completing 2 assessments, 18.3% (n=1,962) completing 3 assessments, 9.7% (n=1,038) completing 4 assessments, and 9.9% (n=1,057) completing 5 assessments. The mean baseline PHQ was 13.36 (SD 4.96) and did not significantly vary with the total number of assessments completed. The mean end PHQ was 10.80 (SD 5.83) and was significantly lower the more assessments were completed. Patients participated in chat conversations throughout the study period, as well as in the three weeks leading up to the baseline assessment in some cases, which were included when available (weeks -3 through -1). Messages were aggregated by concatenating them (i.e. combining messages in sequence), creating a single 'document' as unit of analysis. For studies 1 and 2, each PHQ score was used to label the pooled messages from the period on which the questionnaire asks respondents to reflect (the two previous weeks). PHQ-9 questionnaires were filled out at the beginning of weeks 0, 3, 6, 9, and 12. For study 3, messages were pooled by week (starting with week -3), and each series of (up to) 15 datapoints has one trajectory label. On average, participants had 7.4 weeks of messages and wrote 770 words per week; patients completed 2.2 assessments on average and wrote 2,133 words per completed assessment. At baseline, the number of words written did not vary significantly with depression severity ($p=0.33$). There were 79,096 weeks of messages and 23,950 PHQ assessments with messages. For discussion of the relationship between demographic and engagement factors and treatment outcomes, please see Hull et al. [190].

4.2.3.1 Trajectory labels

Based on patients' longitudinal PHQ-9 and GAD-7 scores, Hull et al. clustered patients using latent growth modeling and assigned the following labels to the six trajectory groups that emerged: Acute Recovery, Recovery, Depression Improvement, Anxiety Improvement, Chronic, and Elevated Chronic. The middle two categories appeared to capture patients who improved in some symptoms but not others. Additionally, improvements in PHQ (or lack thereof) were less clear than in the other groups. Because the individuals in these groups are thus outside the simple definitions of depression "improvement" and "non-improvement", they were not included in analyses of binary improvement status. A subset of 6,760 patients was used for trajectory analysis. Of these patients, which Hull et al. identified as strictly "improving" or "non-improving", 47.2% (n=3,189) improved (classified as Recovery or Acute Recovery), and 52.8% (n=3,571) did not improve (classified as Chronic or Elevated Chronic).

The present work focuses on ascertaining the utility of linguistic markers to predict depression symptom improvement only; therefore, when using trajectories, we simplified trajectories into a binary "improvement" label, with the two recovery classes in the improvement group and the two chronic classes in the non-improvement group.

4.2.4 Statistical analysis

4.2.4.1 Study 1: Association of linguistic markers with PHQ

To validate the basic premise of LIWC and the BA concept, we investigated the relationship between linguistic markers and PHQ-9 scores using the (up to) five measurements of linguistic indicators with corresponding PHQ-9 scores per patient. For this analysis, each pair of PHQ-9 assessment and corresponding message log was treated as a data point. We first determined whether the established LIWC metrics as well as our novel BA metric are statistically significantly different between patients with different depression symptom severity. Severity was defined by the clinical depression symptomatology groups used by the PHQ scoring

system: minimal (PHQ ≤ 4), mild (PHQ = 5-9), moderate (PHQ = 10-14), moderately severe (PHQ = 15-19), and severe (PHQ ≥ 20). Further, the average difference in each linguistic marker for each unit difference in PHQ score was determined using mixed-effects linear regression, treating the patient identity as a random effect.

4.2.4.2 Study 2: Utility of BA subconstructs

Each question of the BADS activation subscale aims to capture a distinct dimension of the theoretical construct. To determine the difference between the components and the potential clinical value of the sub-components compared to pronoun usage, affect measures, and the overall BA concept, we conducted regression analyses on combinations of different variable subsets. For each analysis, we determined the variance explained by each subset of predictors in a mixed-effects model with PHQ-9 score as the outcome, treating participant identity as a random effect. Predictors were combinations of (1) subsets of the established LIWC variables (first-person singular pronouns, first-person plural pronouns, emotional tone, or all three; sadness, health, biology, or all three) and (2) subsets of the behavioral activation variables (the overall construct, each of the seven subconstructs, all seven subconstructs, and all seven subconstructs plus the overall construct). Comparing the amount of variance explained (R2) between baseline models and models that include additional variables yields insights into the extent to which the added variables provide further information. However, chance associations alone can increase R2 even if variables provide little usable additional information; therefore, we additionally determined the Akaike information criterion (AIC), which penalizes model fit in response to model complexity. A non-increased AIC in conjunction with an increased R2 should therefore signal that added variables contained new information.

4.2.4.3 Study 3: Association of linguistic markers with patient trajectories

To determine the association between different linguistic indicators and outcome, mixed-effects linear regression was used to compare the rates of change of the variables over

time between patient trajectories (whether patients were improving, i.e. classified as Recovery or Acute Recovery, or non-improving, i.e. classified as Chronic or Elevated Chronic).

We compared the average change in each variable per 1-week difference (regression slope). For this analysis, messages were aggregated by week, yielding a time series with up to 15 data points for each patient. Thus, we calculated how PHQ scores and linguistic indicators changed with time in the improving and non-improving groups, controlling for the within-patient dependency of samples. Specifically, for each of the two groups, we fitted a mixed-effects linear regression model of the following form:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_{0i} + \gamma_{1i} X_{ij} + \epsilon_{ij}$$

Where Y_{ij} is the variable of interest measured for participant i in week j, e.g. PHQ, activation, or satisfaction. X_{ij} is the week number. $\beta_0$and $\beta_1$are the fixed effect (time) parameters. γ_{0i} and γ_{1i} are the random effects (participant ID) parameters. Calculations were done using the statsmodels package in Python [191].

4.3 Results

4.3.1 Study 1: Relationship between linguistic indicators and PHQ scores

Using LIWC to measure the percentage of words belonging to the overall activation construct (including all terms related to any of the subconstructs), the average level of activation across the baseline chat logs was 3.66 (SD 0.89) and varied significantly with the depression symptom severity category (Figure 4.1), as did the LIWC emotional tone measure and the LIWC pronoun measures (first-person singular and first-person plural). Less depressed individuals used more “we” pronouns and fewer “I” pronouns. All individuals expressed more negative affect than positive (tone < 50), with the most depressed individuals exhibiting an emotional tone balance most extremely tipped toward negative affect (lowest scores). The topic-related word categories were also significantly different between severity groups, with sadness having

the most pronounced differences between groups. The health and biology categories appear to show increased usage in more depressed individuals, but have remarkably large confidence intervals for the least depressed group (none/minimal). Higher overall behavioral activation levels were detected for lower depression levels, indicating that patients with more severe depression symptoms discussed activities and associated feelings of enjoyment and reward less than their less-depressed counterparts.

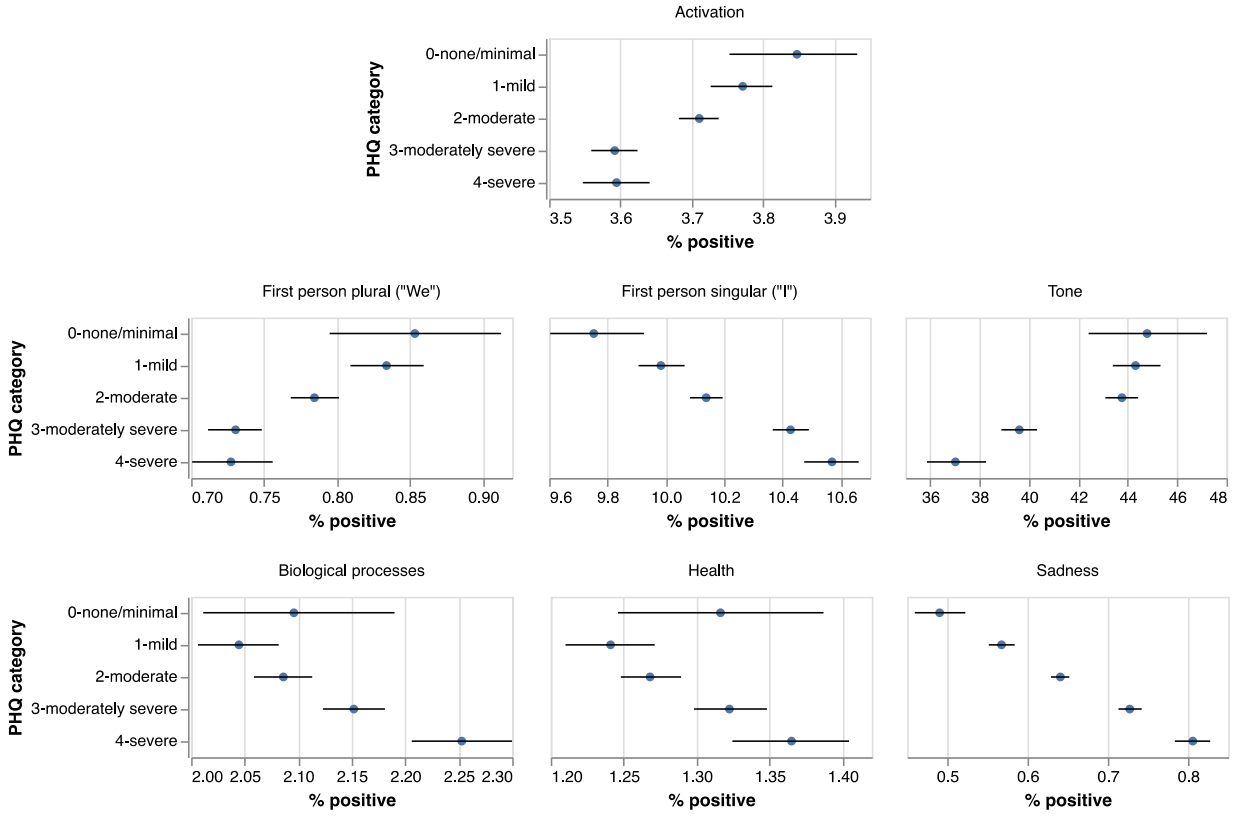

Figure 4.1 Mean of each LIWC measure by depression symptom severity category at baseline. Minimal (PHQ<=4, n=393), mild (PHQ=5-9, n=1,865), moderate (PHQ=10-14, n=4,109), moderately severe (PHQ=15-19, n=3,002), severe (PHQ>=20, n=1,331). Bootstrapped 95% confidence intervals are shown. All variables shown had statistically significantly different means across groups according to one-way ANOVA (p<0.05).

4.3.2 Study 2: Utility of subconstructs

The variance in the overall PHQ score explained (R2) by the 109 models fitted to all possible combinations of the LIWC and BA variable sets is shown in Figure 4.2. The amount of variance explained for each baseline model is shown as bars with strokes, and comparison

values are shown without. Darker colors indicate better fit as measured by the AIC. At baseline, emotional tone is more informative than "I" or "we". The health category is the most informative topic category; the three topics together explain more variance than the tone and pronoun variables together, and also have better fit. All LIWC variables together explain the most variance without detracting from model fit. Of the activation subconstructs, the decision, long-term planning, and daily structure components are most informative; again, all variables together explain the most variance. Including tone added more to satisfaction, breadth, accomplishment, and effort than to decisions, long-term planning, and daily structure. LIWC-package constructs alone (tone, I, we; content topics) accounted for 68.3% of the variance, while the combination of our newly created behavioral activation sub-constructs plus total score alone accounted for 69.7% of the variance. The highest R2 of 80.4% was achieved by the model that included all variables: emotional tone, function words (I, we), all three topic categories, and all seven sub-components of activation, along with the overall activation level. Interestingly, including the overall BA concept along with the BA subconstructs appeared to improve both R2 and fit compared to the subconstructs alone.

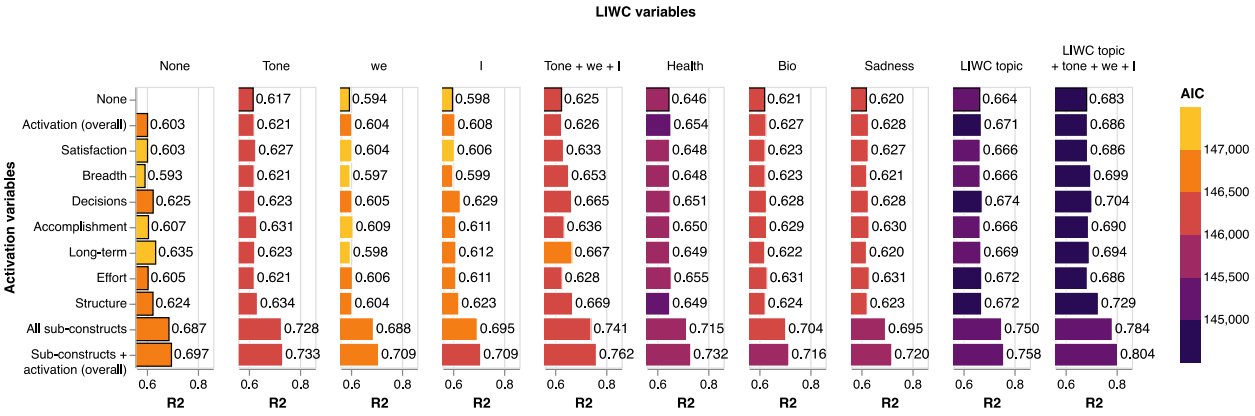

Figure 4.2 Variance explained (R2) by each subset of variables in a mixed-effects model with PHQ score as the outcome.

Baselines are shown with outlines (left column and top row). Compare columns to the first column for the increase in R2 due to standard LIWC variables compared to activation variables alone; compare rows to the first row for the increase in R2 due to activation variables compared to standard LIWC variables alone. Colors indicate the relative Akaike information criterion (darker colors indicate better fit).

4.3.3 Study 3: Relationship with patient trajectories

Figure 4.3 shows the average change in each linguistic marker per week in the improving and non-improving group. Several linguistic indicators showed average amounts of change over time that were significantly different between the two groups.

Of the established LIWC markers, emotional tone, first person singular pronouns, first person plural pronouns, and biology words were different between groups. Interestingly, biology word usage decreased less in the non-improving group than in the improved group, while health word usage decreased more in the non-improving group. Sadness was reduced in both groups over time, with a larger change in the improving group, though the difference between groups was not statistically significant.

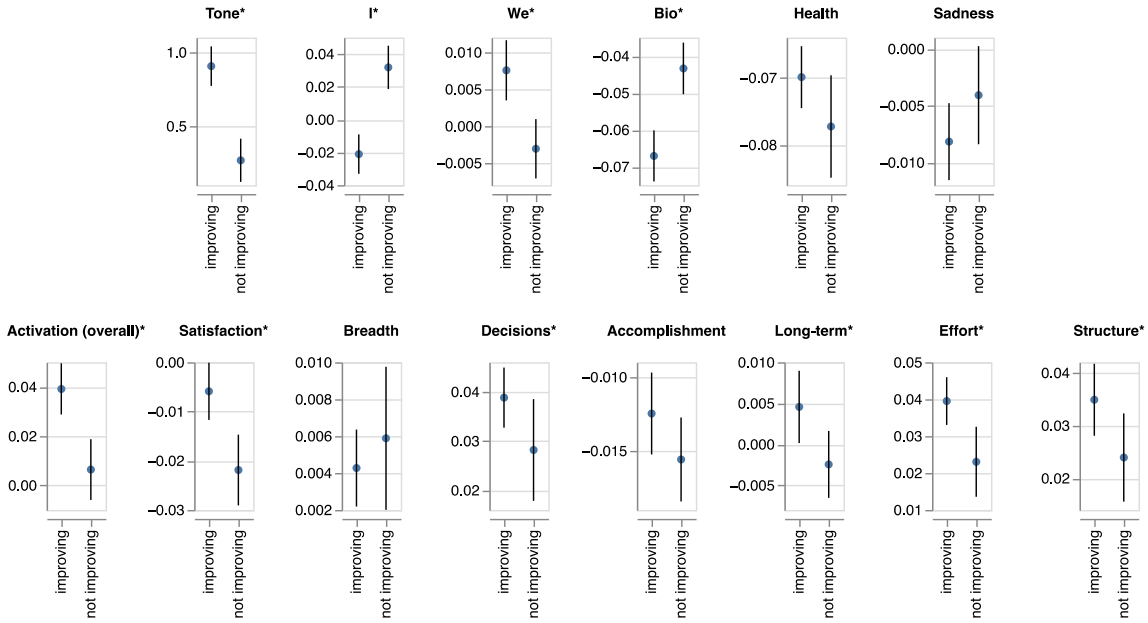

Figure 4.3 Regression coefficients and corresponding 95% confidence intervals of the Mixed Effects models, i.e., the average change in the given variable for each treatment week.

*** = p<0.05**

Of the linguistic markers of behavioral activation, the markers for satisfaction with activities and rewarding effort had the most pronounced difference between groups, along with the overall activation marker. The fitted fixed effects models are shown in Figure 4.4. Neither

the breadth of activities discussed or mentions of feelings of accomplishment were different between groups.

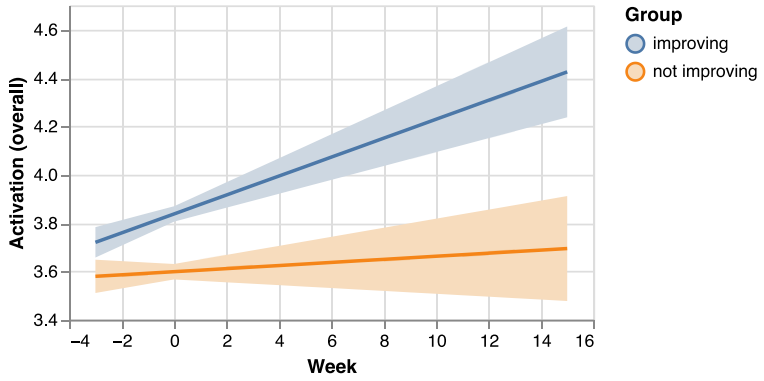

Figure 4.4 Fixed effects of the fitted linear mixed effects models for the improving and non-improving groups for Activation (overall).

Improving: activation = 3.837 + week * 0.039. Non-improving: activation = 3.598 + week * 0.006

4.4 Discussion

In this work, we developed an approach to measuring behavioral activation from patients' natural language and demonstrated that this metric, like scores collected with the validated PHQ-9 depression symptom questionnaire, can discriminate between patients on positive and adverse trajectories. Furthermore, we demonstrated that activation can be measured in terms of its distinct, clinically meaningful sub-constructs, each of which is complementary to established linguistic indicators (affect, pronouns, content categories), and has potentially different uses in future research and applications. Additionally, we demonstrated that established linguistic indicators obtained using LIWC are associated with changes in depression symptomatology as measured with the PHQ in a large sample of online therapy clients.

At baseline, the depression severity experienced by patients in this study was well captured by linguistic markers. The established LIWC markers of affect, pronouns, and topic categories were significantly different between PHQ-9 severity groups. Remarkably, sadness and the indicator of self-focus show nearly identical patterns across groups. Though depression

without sadness does occur, sadness is perhaps most famous marker of depression, and the agreement between first person singular pronouns and the marker of sadness confirms the theory of self-focused attention and provides further support for the measurement of self-focus using LIWC. Further, even at baseline, BA is markedly different between depression severity groups. The presence of this association at baseline provides support for the connection between activation and depression symptoms on which BA therapies are based.

Notably, not all variables are able to differentiate all depression severity groups. For example, tone appears to be clearly different between the moderately severe and severe groups at baseline, while activation cannot differentiate between the two. Conversely, mean activation is different between moderately depressed individuals and those with only minimal symptoms, while tone was not significantly different between these groups. That not all variables differentiate may be related to the polymorphous clinical presentation of depression; patients experience different combinations of symptoms to varying degrees [192]. The heterogeneous presentation of depression suggestions that any single metric can only partially capture the clinical severity of depression and underscores the importance of measuring multiple aspects of depressive symptomatology. Combining several complementary markers, each carrying some unique information, can maximize predictive power by capturing different dimensions of depressive psychopathology.

The mechanisms underlying improvement due to BA therapies are not sufficiently well understood. A comparison of different BA approaches found that most result in similar benefits, even though they include slightly different protocols, suggesting that some elements of the various activation approaches may be unnecessary [167]. By investigating the individual components of the activation subscale of the BADS in isolation, we showed that each contributes further information: Considering the individual parts, rather than just the overall idea of activation, explains more variance in regression models without deteriorating model fit, which might be expected if more variance was explained merely because more variables were included.

This finding suggests that the features capture distinct dimensions of activation and that some dimensions may be more important than others. For example, we found that the idea of breadth (corresponding to BADS item "I engaged in a wide and diverse array of activities.") was the least predictive of patient improvement over time. Manos et al. previously showed factor analysis that a modified version of this item ("I engaged in many different activities") should be retained in a short version of the BADS [193]. However, the current work interpreted the item as capturing activity diversity specifically, counting ideas such as 'exercise', 'restaurant', and 'instrument', similar to the Pleasant Events Schedule (PES) questionnaire. The PES asks respondents for the number of times they participated in activities on a list of 320 activities in the past 30 days, including "Playing baseball or softball", "Going to a restaurant", and "Playing a musical instrument" [194]. Manos et al. point out that PES may not represent the key functional activities of every respondent and therefore may not accurately capture activation, and their rewording of the item removes some of the focus on activity diversity. Our findings support the idea that focusing on activity diversity may not be beneficial for the identification of linguistic markers of behavioral activation. Future work might modify the dictionary created here to remove the specific named activities. Further research may also grant insights into opportunities to refine treatment interventions and protocols. For example, it may be prudent to explore modifying BA therapies to remove any emphasis on activity diversity and focus on frequency of patients' activities of choice.

Our analysis shows that emotional tone, "I", and "we" are strongly associated with patient trajectory, consistent with prior work. Both the improving and non-improving groups shifted their tone to be more positive over the treatment period (positive, non-zero slope), which may be an effect of participating in therapy – e.g. participants may be focusing on solutions rather than problems in conversations with their therapists. While the sadness category was also strongly associated with patient trajectories, the effect was less pronounced than may have been expected, considering that feelings of sadness are perhaps the most common symptom of

depression. Surprisingly, the health topic decreased in usage more in those not on a positive trajectory than those who experienced improvement, while the "biological processes" topic group decreased more in the improving group. The biological processes topic group contains health words, as well as ingestion, sexual, and body words. That patients on the path to recovery discuss these topics less as time passes may indicate that they experience fewer somatic symptoms.

In addition to the overall BA construct, we found several sub-constructs that were strongly associated with patient trajectory, i.e. satisfaction, decisions, long-term planning, effort, and structure all showed significant differences between the improving and non-improving group ($p<0.05$). The breadth of activities and feelings of accomplishment were the most similar between these groups out of all activation sub-constructs. However, the addition of breadth to the established LIWC markers resulted in more variance explained without a reduction in goodness of fit. In fact, breadth was the third most informative single activation construct when included alongside all topic and pronoun variables and tone. In other words, while breadth alone is not explanatory, it is informative in the context of other predictors such as emotional affect. This observation may indicate that discussing activities without positive feelings is not indicative of improvement; activities must also be perceived as rewarding.

Pennebaker et al. first began developing and using LIWC in the 1990s [122]. Counting words belonging to semantic and syntactic categories is simple yet effective, as demonstrated in countless experiments across several fields [122]. For example, LIWC has been used to compare linguistic style and content between essays written by depressed and non-depressed students [172]. In this study, depressed students exhibited increased use of first-person singular pronouns and negative affect. De Choudhury et al. used LIWC categories, including emotional tone and linguistic style, and other indicators to predict vulnerability to depression from another passively collected record of naturally occurring digital discourse, namely social media data [123,195]. Their findings provide further evidence that both affect and linguistic style are

predictive of current and future depression. LIWC has also been used in the clinical setting. Sonnenschein et al. [133] compared linguistic markers in transcripts from in-person psychotherapy session between patients with depression (without anxiety) and anxiety (without depression) and found a significant difference in emotion expression, but only minor differences in pronoun usage; the study did not include healthy controls. Molendijk et al. [196] found significant differences in pronoun usage as well as negative emotion between essays written by psychiatric and non-psychiatric patients, but found that the effect was not specific to depression. A detailed review of LIWC applications is available elsewhere [122].Despite this extensive existing work utilizing LIWC, sample sizes have historically been small: the largest sample size in Edwards et al.'s meta-analysis of first person singular pronoun use in depression was 966 [173]. Containing chat conversations from over 10,000 individuals over almost three years, with over 74 million words, our dataset of naturally occurring language is considerably larger than any used to validate the relationship between LIWC variables and depression in previous experiments. Our results provided further validation of LIWC's tone and pronoun related variables, in the context of scores from a validated, standardized instrument for measurement of symptomatology. Changes in established LIWC variables are consistent with case-control differences demonstrated in prior research [123,172,177], with tone and first-person plural pronouns increasing as symptoms decrease, and first-person singular pronoun usage decreasing with improvement in symptoms. Therefore, we showed that these metrics indeed reflect depression symptom status in this dataset, providing further strong support for their relationship to depression. However, in the context of the current data, BA variables explained more of the variance in the PHQ data than pre-existing LIWC constructs.

4.4.1 Limitations

Reliance on the PHQ-9 is a limitation because this instrument focuses exclusively on symptoms of depression. Future studies may use instruments assessing wellbeing and social adjustment.

The trajectories used to categorize patients as improving or otherwise were assigned via unsupervised learning and do not directly correspond to the total change in PHQ over the course of treatment. Rather, they account for the entire series of depression and anxiety scores over the treatment period. A participant may have an absolute decrease in PHQ score (e.g. 16, 18, 15, corresponding to a total change in PHQ of one point between the beginning and end of treatment); however, the patient may not be experiencing clinically meaningful progress. Thus, depending on the entirety of the PHQ-9 as well as GAD-7 scores over the treatment period, the model may not assign an "improvement" category for such a patient. Many patients only had one or two scores, and some participants with identical sets of scores were assigned to different groups. However, mixed-effects linear regression confirmed that depression symptoms as measured by PHQ-9 scores improve significantly more for patients in the improving group than for patients in the non-improving group. Thus, we believe that the categories are sufficiently accurate for our purposes. While we excluded the "gray area" trajectories of Depression Improvement and Anxiety Improvement, it worth noting that analyses were repeated with these included, and results, though less clear, were not different and our conclusions held; this is the expected result of introducing additional noise. Another potential limitation of the trajectories is that even the "non-improving" group showed a slope in PHQ that was significantly different from 0; we believe that this is an effect of unsupervised learning (i.e. the trajectories clustered into groups labeled as 'chronic' are distinct from the other clusters, but not necessarily entirely without improvement) coupled with the fact that therapy and just the simple passage of time (regression to the mean) is at least somewhat effective for most people. A truly "non-improving" control group is therefore difficult to carve out in this dataset, possibly because such people

represent a minority of participants and were thus not assigned a separate cluster by the unsupervised approach. Future work on this dataset may consider using the more granular patient trajectories if the degree of improvement or exacerbation is of interest. For example, a patient may be responding well ("remission"), but there may be room for improvement ("acute remission"); alerting providers to this situation could result in additional efficiency of care.

The lexicon of words collectively representing activation developed in this work was based solely on the messages in this data set. This approach ensured that the concept represents language usage in our specific study population. While domain-specific texts generally work better for distributional semantics than more general corpora, they may also result in artifacts with limited generalizability. Our study population was predominantly young (55% were between 26 and 35 years old) and, considering that they used a paid online therapy service, presumably financially stable. 28.5% were residents of California or New York, and 78.9% were female. This striking gender imbalance is consistent with previously reported gender imbalances in both online and face-to-face therapy. For example, Chester et al. [197] report that 70% of online therapy clients are female and a recent comparable online mental health service in Australia [198] reportedly had 72% female clients. Sagar-Ouriaghli et al. [199] report that women are 1.6 times more likely to receive any form of mental health treatment than men. Nevertheless, the makeup of our study population must be considered in future applications of our results. Geographically and demographically diverse groups may significantly differ in their word choice and usage, and as a result, the lexicon may not be generalizable to other groups.

Word count approaches such as the one employed by LIWC have limitations. While most modern NLP methods account for negations, counting words does not. LIWC has a separate "negation" category, but it only considers single words, and thus does not assign negation statuses to individual concepts. For example, describing having planned the day's activities and *not* having planned the day's activities would both count toward our *structure* concept in the same way. Because both statements reflect that the patient engaged with the idea of planning

their day's activities, this is arguably a minor limitation. Still, our results indicate that context is more important for some concepts than others; for example, breadth is only informative when considered alongside other markers such as emotional affect to contextualize it. In our future work, we plan on utilizing approaches that can account for negations.

A potential limitation is that the therapy sessions in this work did not specifically use BA therapy. However, behavioral activation is a common pathway of many therapies, which through a variety of interventions increase exposure of depressed patients to meaningful, rewarding experiences.

An important factor to consider when proposing such automated analyses is the degree to which passive monitoring of therapist-client communication for the purpose of measuring BA may be construed as invasive or intrusive. While we did not directly engage with patients in the current work, we note that our recent work in the suicide prevention domain provides some indication that passive monitoring of this sort may be acceptable to patients when conducted by a trusted party, with 68% of survey participants indicating that the automated analysis of personalized web search data for suicide prevention would be acceptable provided this triggered minimally invasive interventions (such as connection to a support network or therapist) only [94].

Another limitation of the study is the absence of qualitative review that could examine the accuracy of our ratings.

4.4.2 Future work

An additional implication of the potentially causative mechanism of activation is that it should occur before symptom improvement. While more direct measures of sentiment and mindset reflect an individual's current thoughts and feelings, "activity" topic analysis may reveal long-term dimensions of patient trajectory. Discussing activities, plans for activities, or even avoidance of activities shows that a patient engages with the ideas surrounding BA and may

indicate that a patient is moving toward or already part of a positive feedback loop of self-perpetuating improvements. Consequently, one may expect metrics of activation to predict longer-term changes more accurately than word use analysis, which may be confounded by the mood of the moment, and thus capture a separate and clinically meaningful dimension of symptomatology and treatment success.

Our results show that the BA sub-constructs of *decisions* (independent decision-making), *long-term* (planning for the long term and acting accordingly), and *structure* (structuring daily activities) are parallel in terms of the amount of variance they explain. Let us call these *activities*. Similarly, *satisfaction*, *accomplishment* (a sense of having accomplished something), and *effort* (enjoyment due to effort exerted) are similar to each other. Let us call these *reactions*. Activities alone were more explanatory of improvement than reactions alone; in other words, activities are explanatory even without considering emotional tone. Adding tone increased the amount of variance explained for reactions to the levels for activities alone. In other words, reactions are more informative when considered alongside tone than without, whereas adding tone makes comparatively little difference for activities. A possible explanation is that merely discussing feelings of satisfaction, or lack thereof, may not correspond to improvement. However, engaging in long-term planning, scheduling activities, and structuring routines does correspond to symptom improvement. Mentions of being satisfied may only be triggered positively once the positive feedback loop is set in motion. Conversely, this pattern is in agreement with the helplessness theory [174,200], which states that experiencing negative consequences to reward-seeking behavior perpetuates the avoidance phenomenon, and therefore continuing depression symptoms, by supporting a negative outlook on the future. Therefore, it requires additional information about the patient's tone to be informative of current symptom severity. Therefore, an opportunity for future work is to investigate the temporal relationship between symptom improvement and the constructs in the activities group compared to those in the reactions group. One may be more useful than the other to predict

long-term changes. Ascertaining the timeline of changes in these different activation metrics relative to patient trajectories is a prerequisite to the translation of linguistic indicators into clinical insights.

Recently, a broad range of sophisticated natural language processing techniques have gained traction. Our work here demonstrated that theme and affect analysis conducted via established, straightforward methods such as word counting both reflect and potentially predict depression status. It is plausible that more advanced methods may surpass those used here. Consequently, future work is set to include a deep neural network model trained on GoEmotions [51], a large corpus of social media posts annotated for the extent to which they express a set of emotion categories [201].

Having validated the linguistic indicators of depression and activation presented here, the question of how to incorporate them effectively into care processes remains. Predictive analytics solutions must be operationalized effectively to improve health outcomes. Accordingly, future work should focus on testing and iteratively refining these measures as part of care delivery. For example, metrics of depression symptom severity and behavioral activation may be automatically extracted from patient messages on virtual chat therapy platforms and used to guide the therapist in recognizing whether a patient responds well to therapy or if a change in direction is appropriate.

4.5 Conclusion

This work makes several key contributions. First, we devised a computational method to automatically assess theoretical constructs of BA from patient language. Second, building on prior work demonstrating that activation has a close relationship with depression scores, we demonstrated this new metric reflects depression symptom severity. Third, we validated established linguistic markers of depression in a large corpus of naturally occurring language collected as part of psychotherapy sessions, presenting differences between participants with

low and high PHQ-9 scores. Fourth, we showed that both the well-established LIWC measures as well as the novel BA measures have utility in predicting longitudinal patient trajectories. Finally, we demonstrated that our metrics of the individual subconstructs of BA capture distinct dimensions of the underlying mechanisms and may lend themselves to unique clinical insights. This work therefore enables further work in automated diagnosis and assessment of depression, as well as refinement of BA psychotherapeutic strategies.

4.6 Acknowledgements

This work supported by the National Library of Medicine [grant number 67-3780] (HAB), the National Institutes of Mental Health [grant numbers R01 MH102252, P50 MH113838] (GSA), [grant number P50 MH115837] (MDP, PAA).

Chapter 5. Comparing emotion feature extraction approaches for predicting depression and anxiety

Aim 1 (Chapter 3) of this dissertation revealed the need to take patients' unique situation into account when providing support for the purpose of suicide prevention. Current and historical emotional state are important indicators not only of patient well-being, but also of ways to best support the patient. For example, a patient who has been experiencing grief due to the loss of a loved one may benefit from educational materials for coping with grief; on the other hand, a patient who has a tendency to react to difficult situations with anger may benefit from emotion regulation strategies. However, comprehensively reviewing the history of messages exchanged between intervention staff and the patient for patterns of emotional expression can be time intensive, suggesting an opportunity for automated emotion extraction.

In the work described in this chapter, my co-authors and I investigated two different approaches to extracting emotion features from text written by patients as part of asynchronous, message-based therapy. Similar to the work described in aim 2 (Chapter 6), this work benefitted from social media data: GoEmotions, the neural emotion extraction approach used here, was trained on Reddit posts annotated for 27 fine-grained emotions. I demonstrated the ability to use these two approaches to extract emotion from text messages exchanged between clients and their therapists during the course of treatment, and found that extracted features were associated with depression and anxiety symptom severity.

This work is also uniquely relevant to informatics-support suicide prevention. The emotion feature extraction approaches investigated here could be used to extract emotions from current and historical messages exchanged in the Caring Contacts intervention. This may inform how intervention staff interact with a patient to validate and support them, or what materials clinicians share with the patient to help them bolster their coping strategies.

A version of this chapter was previously published as part of the Association for Computational Linguistics (ACL) under a Creative Commons Attribution 4.0 International License. © the authors.

Burkhardt H, Pullmann M, Hull T, Areán P, Cohen T. Comparing emotion feature extraction approaches for predicting depression and anxiety. In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics; 2022:105-115. doi:10.18653/v1/2022.clpsych-1.9

Abstract. The increasing adoption of message-based behavioral therapy enables new approaches to assessing mental health using linguistic analysis of patient-generated text. Word counting approaches have demonstrated utility for linguistic feature extraction, but deep learning methods hold additional promise given recent advances in this area. We evaluated the utility of emotion features extracted using a BERT-based model in comparison to emotions extracted using word counts as predictors of symptom severity in a large set of messages from text-based therapy sessions involving over 6,500 unique patients, accompanied by data from repeatedly administered symptom scale measurements. BERT-based emotion features explained more variance in regression models of symptom severity, and improved predictive modeling of scale-derived diagnostic categories. However, LIWC categories that are not directly related to emotions provided valuable and complementary information for modeling of symptom severity, indicating a role for both approaches in inferring the mental states underlying patient-generated language.

5.1 Introduction

Almost 10% of adults in the United States receive mental health counseling [202]. The principle of measurement-based care dictates that medical treatments should be initiated and

evaluated over time based on repeated assessments of patient symptoms and symptom trajectory [163]. In the context of talk therapy, mental health practitioners estimate treatment progress based on patients' current and historical verbal communications. For evaluating depression and anxiety severity, expressions of emotional state are key aspects of such communications [178,203,204].

While prior work predominantly focused on sentiment, i.e. positive/negative polarity, expression of fine-grained emotions [52,53] may give further insights into depression and anxiety symptomatology. For example, pride may be impacted by depression in a unique way. Gruber et al. [54] showed that pride, a positive emotion relating to the self, is inversely correlated with depression, which is often associated with a poor self-image. At the same time, they found a smaller effect on joy and amusement, concluding that grouping these emotions into "positive affect" may result in a loss of nuance.

The increasing adoption of digital mental health tools and services, particularly message-based therapy, has afforded new opportunities to assist practitioners in quantifying depression and anxiety severity by assessing emotion in patient-generated text. Linguistic Inquiry and Word Count (LIWC) [121,122] is a software package designed to count words belonging to pre-defined categories with an extensive track record of validation for the detection of linguistic indicators of mental state [122]. It is commonly used to measure positive and negative affect, a limited set of specific emotions (sadness, anxiety, and anger), and other linguistic dimensions related to style and topic. Several LIWC categories have established relationships with depression, including the affect category sadness (e.g. "sad", "cry", "suffer"), the topic category health (e.g. "alcohol", "rash", "self-care"), and the syntactic category first-person pronouns (e.g. "I", "me", "my"). LIWC has been used to measure depression levels in social media posts [123,195,205,206], therapy conversations [133,135], and other written texts [172,177]. LIWC measurements have also been shown to distinguish between patients with depression and those with anxiety disorders [133], correlate with self-reported measures of anxiety and worry in

written descriptions of emotional responses to COVID-19 [207], and predict whether posts emanated from anxiety-related subreddits [208].

However, word counting methods cannot address linguistic phenomena such as negation ("not bad"), sarcasm, and context-dependence (for example, in the case of polysemy, words have multiple meanings that can only be disambiguated in context), and manually defined dictionaries may omit synonyms for terms they encode. Prior work suggests that neural network (NN)-based natural language processing (NLP) techniques can account for such phenomena and may therefore improve upon this straightforward word-counting method in their ability to identify concepts related to symptom severity. Shen and Rudzicz found that the performance of machine learning models identifying whether or not Reddit posts were drawn from anxiety-related subreddits improved when these models included neural word embeddings rather than LIWC-derived features [208]. However, the distributed representations of posts used in this work do not relate directly to interpretable emotion features. Further, contemporary transformer-based NN language models offer advantages over neural word embeddings in their ability to leverage proximal cues (such as "not") when interpreting the contextual meaning of a word. As noted by the authors, this work suggests a need for further research on automated assessments of linguistic indicators of anxiety disorders, involving larger data sets and explicit diagnostic assessments.

Therefore, using a large set of messages from text-based therapy session, we investigated if emotions extracted using a Bidirectional Encoder Representations from Transformers (BERT) [100] based model trained on GoEmotions, a large dataset of Reddit posts annotated with 27 fine-grained emotions [51], are stronger predictors of depression and anxiety status than counts of emotion-related word categories (LIWC). To this end, we first determined the association of each feature with the outcomes of interest in univariate regression analyses. Further, in order to provide clinical decision support to mental health practitioners, it is paramount to be able to classify previously unseen messages as indicating depression and/or anxiety. We therefore

proceeded to train and evaluate a machine learning classifier using emotion features in conjunction with established depression-related LIWC features to predict depression and anxiety status in a held-out test set.

5.2 Methods

5.2.1 Data

We utilized a corpus of messaging therapy sessions from over 6,500 unique patients previously collected via the Talkspace platform [190]. Talkspace offers a paid service utilizing licensed and credentialed therapists to conduct asynchronous, message-based therapy conversations. All patients and clinicians give written consent to the use of their data in a de-identified, aggregate format as part of the user agreement before they begin using the platform. Over the course of 12 weeks, patients engaged in two-way messaging therapy and completed depression questionnaires (9-item Patient Health Questionnaire, PHQ-9 [165]) as well as anxiety questionnaires (7-item General Anxiety Disorder questionnaire), every 3 weeks. For each available score, patient messages from the period in question ("(o)ver the last two (2) weeks") were concatenated into a single unit of analysis ("document"), resulting in up to 4 labeled data points per patient (weeks 3, 6, 9, and 12). All messages without a corresponding score were excluded from analysis. Data from baseline assessments were removed, as preliminary analysis suggested that messages before the week 0 mark introduced spurious associations due to differences between typical therapy dialog and the patient-therapist matching process, combined with generally worse symptom severity scores at the beginning of the study period. Participants were young (79% were 35 years old or younger), educated (75% had a Bachelor's degree or higher), and predominantly female (79%). Race and ethnicity were not systematically collected. There were over 13,000 text documents with both PHQ-9 and GAD-7 scores, totaling over 24 million words from over 337,000 messages. The original study was approved as exempt by the local institutional review board. The current study concerned

secondary analysis of previously collected de-identified data, which is not considered human subjects research; nonetheless, data were stored on a secure server with study team member access only. All textual data were thoroughly de-identified by an automated algorithm before leaving their source, with all names, places, contact information, social media identifiers, and mentions of specific events removed.

LIWC 2015 was used to obtain the following word-count-based features: first-person singular pronouns ("I"), first-person plural pronouns ("we"), bio, health, sadness, anxiety, anger, positive emotion, and negative emotion. These features were selected on account of their track record of correlation with indicators of depression and anxiety in previous work [122].

A BERT-based GoEmotions classifier pipeline using fine-tuned models available from the Hugging Face transformer library[1] was used to extract emotion features from each document. This model has been shown to approximate published results for performance in extracting emotions from the GoEmotions dataset (macro-average F1 score of ≈0.5 to ≈0.7, depending on the granularity of the emotions concerned). For further details of the training corpus and procedures used, we refer the reader to Demszky et al. [51]. After splitting documents into sentences and extracting emotions from the first 512 tokens of each sentence, scores were averaged over all sentences in a document to yield one set of emotion scores for the two-week period concerned. Only 38 of ≈13,000 documents contained sentences that were truncated due to being over 512 tokens long. The pipeline provides several output settings, resulting in different sets of emotions being extracted. Two sets of emotions were extracted. First, we extracted the set of 6 basic emotions proposed by Ekman [209], consisting of sadness, joy, surprise, disgust, anger, fear, and a neutral category, which was assigned by annotators when they felt that no particular emotion was expressed. Second, we extracted the full set of 28 categories that were used to annotate the GoEmotions corpus, consisting of 27 fine-grained

[1] https://github.com/monologg/GoEmotions-pytorch

emotions described by Cowen and Keltner [201], plus a neutral category. Finally, we calculated positive and negative emotion features by averaging the scores belonging to positive and negative emotions. The negative GoEmotions Ekman emotions are anger, disgust, fear, and sadness; joy is the only positive Ekman emotion. Negative fine-grained GoEmotions (Cowen) emotions encompass anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, and sadness. Admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, and relief are the positive emotions in the fine-grained GoEmotions set. The interested reader is referred to Demszky et al. [51] for further details on these groupings.

5.2.2 Comparison of variables

A common approach to identifying associations of individual variables with an outcome of interest is to determine the statistical significance of the association between each candidate variable and the outcome by fitting univariate regressions. Linear regression models, however, require observations to be independent of each other. Because patients contribute between 1 and 4 observations in our dataset, this independence assumption is not met: two observations from the same patient may be expected to be more like each other than two observations from different patients. Mixed-effect linear regressions can be used to account for this. In such models, the within-patient and between-patient effects of the predictor variables on the outcome are separately accounted for. In other words, in addition to the “fixed effect” of the predictor variables on the outcome (the effect of interest), we model a “random effect” that is different for each patient, which is arbitrary but consistent across all observations for a given patient. In essence, the outcome is the linear combination of an emotion’s global relationship to PHQ-9/GAD-7 scores and the patient-specific relationship of the emotion on scores (plus an intercept term for each effect as well as a residual error term). The univariate mixed-effect linear

regression models for each emotion variable model the patient identity as a random effect and are of the following form:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_{0i} + \gamma_{1i} X_{ij} + \epsilon_{ij}$$

Where Y_{ij} is the ith outcome (PHQ-9 score, GAD-7 score) for patient i, X_{ij} is the level of emotion in the jth document written by patient i, β_0 and β_1 are the fixed effect parameters (emotion), and γ_{0i} and γ_{1i} are the random effect parameters (patient ID), and ϵ_{ij} is the residual error for patient i's jth document. Models were fitted via Maximum Likelihood Estimation using the Statsmodels package for Python [191]. Statsmodels calculates p-values using t-tests. We report the explanatory power of each feature as the amount of variance explained (R2).

Following a similar process, we fitted bivariate mixed-effects models using the positive and negative emotion variables from each feature source.

5.2.3 Prediction

Next, using the Scikit-Learn package for Python [210], we trained random forest classifiers to predict binary depression (MDD) and anxiety (GAD) status from 49 features: 7 Ekman emotion categories from GoEmotion, as well as the positive and negative emotion variables calculated from Ekman emotions; 27 fine-grained emotions plus neutral, as well as the positive, and negative emotion variables calculated from the 27 fine-grained emotions; 5 LIWC emotion variables (positive emotion, negative emotion, anxiety, anger, sadness); and 4 LIWC variables with an established relationship to depression (I, we, biology, health) [123,129,133,135,172]. We first trained random forest classifiers using each individual feature set. Then, we trained models using combinations of these feature sets to evaluate their relative contribution (LIWC non-emotion variables combined with each set of emotion variables from the three sources). Then, we trained another random forest classifier on all available features. For this model, relative feature importance was calculated using SHAP [211].

To avoid information leakage due to within-patient effects [212], data were split into training and test sets such that all observations from an individual patient were kept within the same fold. Patients were assigned to the training (80%) and test (20%) populations, resulting in a training set of 4,913 patients (with 10,006 observations) and a test set of 1,638 patients (with 3,321 observations). Average PHQ-9 across all observations did not significantly differ between training and test observations.

Hyperparameters (number of estimators, maximum number of features, maximum tree depth, minimum number of samples for splitting, minimum number of samples per leaf, using or not using bootstrap) were automatically selected (based only on the training data) via 3-fold cross-validation, a process where, for each hyperparameter combination, each of the three folds is held out in turn, while a model is trained on the remaining 2 folds; this way, 3 scores are produced per hyperparameter combination, and their average represents the score for that hyperparameter set. Finally, the hyperparameters that produced the best score are selected, and a final model with those hyperparameters is trained on all training data, then tested on the held-out test set.

A binary prediction target was used to align predictions with the clinical task of classifying a diagnosis as present or absent. A cut-off between 8 and 11 was previously found to have a clinically acceptable tradeoff between sensitivity and specificity when dichotomizing PHQ-9 scores for diagnosis of major depressive disorder (MDD) [213]. Therefore, we considered a PHQ-9 score of 10 or more (depression severity of moderate, moderately severe, or severe) as indicating MDD for the purposes of this work. A PHQ-9 score of 9 or less (depression severity of mild or none) was considered non-depressed. As the GAD-7 has been found to have acceptable properties for identification of generalized anxiety disorder (GAD) at a cutoff of 7-10 [214,215], a GAD-7 score of 10 or more was considered an indicator of GAD, and a score of 9 or less was considered an indicator of a negative diagnosis for this condition.

5.3 Results

5.3.1 Comparison of variables

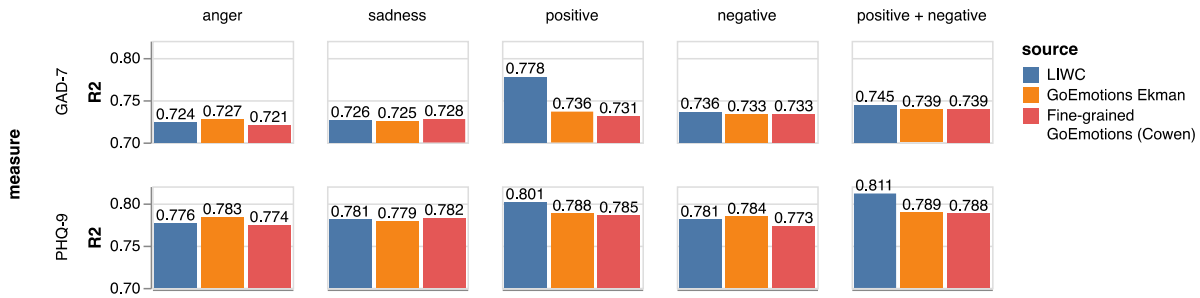

Figure 5.1 PHQ-9 and GAD-7 score variance explained by comparable features from LIWC, GoEmotions (Ekman set), and GoEmotions (fine-grained set)

The variance in PHQ-9 and GAD-7 scores, respectively, explained by each individual emotion variable and by variable pairs is shown in Figure 5.1, Table 5.1, and Table 5.2. Emotion variables that were obtainable from all three feature sources were anger and sadness as well as the summary dimensions of positive and negative emotion. With BERT-based models, these are composites of individual predictions returned by the model, while LIWC returns a summary value as an individual feature. The variance in PHQ-9 scores explained by these directly comparable variables is shown in Figure 1, along with the variance explained by the combination of positive and negative emotion features. The three feature extraction approaches resulted in features that explained similar portions of the variance; LIWC explained slightly more, except for anger and sadness, where the GoEmotions Ekman and GoEmotions Cowan variables explained more, respectively. The GoEmotions Cowan variable for sadness was more explanatory than the GoEmotions Ekman variable, but the Ekman anger variable outperformed the fine-grained anger variable.

Table 5.1 PHQ-9 score univariate mixed-effects linear regression models coefficients and variance explained.

*** p<0.05. ** p<0.001**

GoEmotions		
sadness	18.84 (16.50 - 21.18)**	0.782
admiration	-16.62 (-18.80 - -14.44)**	0.781
annoyance	12.61 (9.67 - 15.55)**	0.778
disappointment	19.01 (16.87 - 21.14)**	0.778
joy	-16.40 (-19.33 - -13.48)**	0.778
pride	-64.35 (-78.54 - -50.16)**	0.777
excitement	-28.34 (-33.18 - -23.49)**	0.777
disapproval	16.11 (12.99 - 19.23)**	0.776
approval	-7.81 (-9.27 - -6.36)**	0.776
confusion	9.65 (7.30 - 11.99)**	0.775
relief	-24.19 (-30.79 - -17.59)**	0.774
neutral	-0.83 (-1.60 - -0.06)*	0.774
anger	18.67 (14.38 - 22.97)**	0.774
disgust	29.79 (21.72 - 37.86)**	0.774
optimism	-6.15 (-8.28 - -4.03)**	0.773
realization	-1.08 (-2.74 - 0.59)	0.773
amusement	-10.96 (-14.85 - -7.07)**	0.772
fear	10.75 (7.44 - 14.06)**	0.771
nervousness	3.44 (0.84 - 6.05)*	0.771
caring	-2.77 (-6.03 - 0.49)	0.771
gratitude	-2.87 (-9.79 - 4.05)	0.771
embarrassment	11.85 (4.25 - 19.45)*	0.771
curiosity	0.03 (-2.56 - 2.62)	0.771
desire	2.08 (-1.10 - 5.26)	0.771
love	-1.96 (-5.22 - 1.31)	0.771
surprise	-4.00 (-10.18 - 2.18)	0.771
grief	134.76 (104.49 - 165.03)**	0.770
GoEmotions Ekman		
joy	-9.31 (-10.21 - -8.41)**	0.788
anger	18.53 (16.46 - 20.61)**	0.783
sadness	15.81 (14.06 - 17.56)**	0.779
disgust	48.43 (37.93 - 58.93)**	0.778
neutral	-0.11 (-1.17 - 0.96)	0.775
surprise	4.11 (2.47 - 5.75)**	0.774
fear	4.52 (2.25 - 6.80)**	0.772
LIWC		
sad	1.21 (1.02 - 1.40)**	0.781
i	0.25 (0.21 - 0.29)**	0.777
anger	0.84 (0.65 - 1.02)**	0.776
health	0.66 (0.52 - 0.80)**	0.775
anx	0.19 (0.05 - 0.34)*	0.774
we	-0.53 (-0.65 - -0.41)**	0.774
bio	0.41 (0.33 - 0.50)**	0.774

Table 5.2 GAD-7 score univariate mixed-effects linear regression models coefficients and variance explained.

*** p<0.05. ** p<0.001**

GoEmotions		
sadness	15.04 (12.96 - 17.12)**	0.728
admiration	-15.02 (-16.97 - -13.07)**	0.727
neutral	-1.00 (-1.72 - -0.29)*	0.725
joy	-16.99 (-19.53 - -14.44)**	0.724
approval	-6.80 (-8.14 - -5.47)**	0.724
fear	18.62 (15.32 - 21.93)**	0.724
annoyance	12.83 (10.20 - 15.46)**	0.724
excitement	-22.74 (-26.98 - -18.49)**	0.723
pride	-56.42 (-69.75 - -43.09)**	0.723
disappointment	14.05 (12.12 - 15.97)**	0.723
disapproval	12.97 (10.18 - 15.76)**	0.723
nervousness	11.91 (9.36 - 14.46)**	0.723
confusion	8.41 (6.33 - 10.48)**	0.721
anger	19.29 (15.38 - 23.19)**	0.721
relief	-22.16 (-28.05 - -16.28)**	0.720
optimism	-6.86 (-8.84 - -4.89)**	0.719
realization	-1.48 (-2.99 - 0.02)	0.718
amusement	-10.34 (-13.73 - -6.96)**	0.717
curiosity	-0.00 (-2.44 - 2.43)	0.717
caring	-1.94 (-5.11 - 1.24)	0.716
gratitude	-3.54 (-7.67 - 0.59)	0.716
desire	1.25 (-1.55 - 4.06)	0.716
love	-3.66 (-7.08 - -0.23)*	0.716
surprise	-6.42 (-11.87 - -0.97)*	0.716
embarrassment	10.33 (3.08 - 17.58)*	0.716
grief	118.01 (90.79 - 145.22)**	0.716
disgust	25.72 (18.68 - 32.76)**	0.715
GoEmotions Ekman		
joy	-8.62 (-9.42 - -7.82)**	0.736
anger	15.92 (14.08 - 17.76)**	0.727
disgust	44.78 (35.38 - 54.18)**	0.726
sadness	12.38 (10.83 - 13.93)**	0.725
neutral	-0.21 (-1.18 - 0.76)	0.722
fear	12.15 (9.97 - 14.33)**	0.722
surprise	3.27 (1.79 - 4.75)**	0.720
LIWC		
anx	0.73 (0.59 - 0.86)**	0.729
i	0.17 (0.13 - 0.21)**	0.726
sad	0.89 (0.72 - 1.05)**	0.726
anger	0.93 (0.76 - 1.10)**	0.724
we	-0.36 (-0.47 - -0.25)**	0.723
health	0.46 (0.33 - 0.59)**	0.717
bio	0.28 (0.21 - 0.36)**	0.716

The combination of positive and negative emotion explained more variance than either positive or negative emotion alone, except when LIWC positive emotion was assessed for GAD-7. Notably, LIWC's positive emotion variable appears to be more explanatory than anger, sadness, and negative emotion for both PHQ-9 and GAD-7, and even the combination of positive and negative emotion for GAD-7.

All individual emotions, as quantified by each of the three feature extraction approaches, are shown in Table 5.1. Expressions of realization, caring, gratitude, curiosity and desire were not significantly associated with either anxiety or depression. Love and surprise were not predictive of depression, but were associated with anxiety. Both were significantly associated with sadness, fear, and nervousness; however, sadness was more strongly related to depression, and fear and nervousness were more strongly related to anxiety. Joy was roughly equally associated with anxiety and depression, across both GoEmotions feature sets. The emotions with the largest differences between more and less depressed individuals were grief, pride, excitement, relief, and disgust. The emotions with the largest differences with respect to anxiety were grief, disapproval, approval, relief, and disgust.

5.3.2 Prediction

In contrast to the multivariate models, results from predictive modeling experiments show a clear advantage for deep learning models, with the best overall performance by ROC and F1 score achieved using GoEmotions Cowen features for both MDD and GAD. As shown in Table 5.3, the models including only the non-emotion LIWC features achieved an area under the receiver-operator characteristic curve (AUROC) of 0.577 for MDD and 0.549 for GAD. When using emotion features only, the fine-grained GoEmotions set performed best. For both MDD and GAD, adding LIWC emotion features to LIWC non-emotion features improved predictive performance less than adding GoEmotions Ekman features, which improved the model less than adding the fine-grained GoEmotions set. Using all emotion features concurrently ("all three")

slightly improved performance for both GAD and MDD (by F1 score but not ROC in the latter case).

The relative importance of all features for the MDD and GAD models is shown in Table 5.4. Fear was ranked higher for predicting GAD than for predicting MDD. Sadness was ranked higher for predicting MDD than for predicting GAD.

5.4 Discussion

In this work, we showed that neural network models such as the BERT-based GoEmotions classifier can outperform LIWC, a straightforward, broadly adopted word-counting method for extracting emotion features from natural language. We further confirmed that some emotions not traditionally associated with depression and anxiety can be predictive of these diagnoses; specifically, pride. Finally, we showed that using LIWC features together with emotion features derived using GoEmotions predict depression/anxiety status with reasonable accuracy. This finding is important, in that further development of such tools could lead to better detection of emotional change during treatment in a way that could be derived naturally in the client/clinician encounter. NLP applied to such naturalistic data has been used for measuring clinician skills in delivering psychotherapy with some success [216]; here, rather than using such tools for quality measurement, linguistic analysis of affect could be used to detect depression/anxiety severity and client response to treatment.

Both LIWC variables and GoEmotions variables explained a large portion of the variance in univariate mixed-effect regressions: R2 values ranged from 0.770 to 0.788 when modeling PHQ-9 scores as outcome, and from 0.715 to 0.736 when modeling GAD-7 scores as outcome. Therefore, LIWC and GoEmotions features both capture valuable information. GoEmotions features marginally outperformed 2 out of 4 of the equivalent LIWC features for predicting GAD-7 and 3 out of 4 features for predicting PHQ-9. For predicting binary depression (MDD)

and anxiety (GAD) status, the emotion set resulting in the best predictive performance when combined with LIWC's non-emotion features was the full GoEmotions set.

Table 5.3 AUROCs, F1 score (positive class), precision, and recall of random forest model trained with just the non-emotion LIWC features, and trained with the non-emotion LIWC features plus LIWC emotion, GoEmotion Ekman and the full GoEmotion feature sets, for predicting MDD (PHQ-9 score ≥ 10) and GAD (GAD7 score ≥ 10).

	AUROC	F1	Precision	Recall
		MDD		
LIWC non-emo	0.577	0.413	0.525	0.341
LIWC emo	0.621	0.471	0.561	0.405
GoEmo Ekman	0.643	0.493	0.583	0.427
GoEmo	0.662	0.522	0.613	0.455
LIWC non-emo +				
LIWC emo	0.640	0.484	0.569	0.420
GoEmo Ekman	0.655	0.498	0.585	0.434
GoEmo	0.671	0.514	0.615	0.441
All three	0.671	0.520	0.612	0.453
		GAD		
LIWC non-emo	0.549	0.290	0.478	0.209
LIWC emo	0.613	0.405	0.541	0.324
GoEmo Ekman	0.643	0.443	0.550	0.371
GoEmo	0.652	0.444	0.565	0.366
LIWC non-emo +				
LIWC emo	0.617	0.401	0.529	0.324
GoEmo Ekman	0.637	0.441	0.548	0.369
GoEmo	0.654	0.451	0.568	0.374
All three	0.657	0.456	0.567	0.382

However, despite the availability of pre-trained models, neural networks can have high computational demands. Consequently, using BERT-based models may not be justified if the cost of model inference outweighs the potential benefits. Therefore, the decision to include these features should be evaluated for each individual predictive analytics project and dataset, weighing the added predictive performance observed at development time with the costs to include the features in production (e.g. a deployed clinical decision support tool continuously evaluating patient-generated messages in real-time), given the available compute resources. Similarly, on-device processing to preserve data privacy can be accomplished with LIWC [217],

but doing this with a BERT-based model would challenge some contemporary and most legacy smartphone devices.

Depression affects individuals in many ways and expresses itself in various behavioral and thought patterns that may not be fully captured with the high-level categories of positive and negative affect. GoEmotions' main strength therefore lies in its ability to extract fine-grained features spanning the breadth of human emotion, capturing depressed individuals' emotional experiences comprehensively. The different emotion feature sets appeared to be somewhat complementary, as evidenced by the additive performance metrics shown in Table 5.3; however, when predicting depression, the combination of non-emotion LIWC features and fine-grained GoEmotions features was as predictive as all features combined, suggesting that all signal is contained within this feature subset. In this work, this breadth enabled us to delineate differences in how different types of emotions are associated with depression and anxiety.

Depression severity was associated with large differences in grief, pride, excitement, relief, and disgust. In agreement with generally lower reactivity [203], less excitement was predictive of depression. Grief manifestations are similar to depression symptoms; though grief in itself is not pathological, it often co-occurs with depression [218]. Additionally, depressed individuals expressing less pride than their non-depressed counterparts might be expected on account of lower self-image, and matches findings presented by Gruber et al. [54]. Caused by a perception of violations of moral and social norms, internally directed disgust, also termed self-disgust or self-loathing, has been reported to be associated with both depression and anxiety symptoms [219]. We further found that increased disapproval - and conversely, decreased approval - were associated with anxiety symptoms. This may be explained by disturbances in interpersonal sensitivity and an inclination to be self-critical, which have been described as characteristic of anxiety [219].

Table 5.4 Random forest classifier features in order of importance (most important first) for predicting MDD and GAD, as calculated by SHAP [211]**.**

GE = GoEmotions. GEE = GoEmotions Ekman

	MDD	GAD
1	LIWC we	GE negemo
2	GEE posemo	GEE negemo
3	GEE joy	GEE joy
4	GEE sadness	GEE posemo
5	GE negemo	GE posemo
6	LIWC bio	LIWC bio
7	GE disappointment	GE fear
8	GEE negemo	GE sadness
9	LIWC sad	LIWC health
10	LIWC i	LIWC we
11	LIWC health	LIWC posemo
12	GE posemo	GE realization
13	GE excitement	GEE sadness
14	GE admiration	GE nervousness
15	GE sadness	GEE fear
16	GEE anger	LIWC negemo
17	GE confusion	GE pride
18	GE pride	LIWC anx
19	GE disapproval	GE joy
20	GE joy	LIWC i
21	GEE disgust	GE disappointment
22	GE realization	GE admiration
23	LIWC posemo	GEE anger
24	GE relief	GE excitement
25	GE approval	GE disgust
26	GE disgust	GE confusion
27	LIWC negemo	GEE disgust
28	GE grief	GE grief
29	GEE fear	GE neutral
30	GEE neutral	GE relief
31	GE fear	GEE neutral
32	GE desire	GEE surprise
33	GE remorse	LIWC sad
34	GE curiosity	GE desire
35	GE nervousness	GE neutremo
36	GE embarrassment	GE curiosity
37	LIWC anx	GE gratitude
38	GE optimism	GE disapproval
39	GE amusement	GE love
40	GE neutremo	GE embarrassment
41	GE neutral	GE anger

42	GE gratitude	GE approval
43	GE love	GE annoyance
44	GE annoyance	GE amusement
45	GE surprise	GE remorse
46	GEE surprise	GE caring
47	LIWC anger	GE surprise
48	GE caring	GE optimism
49	GE anger	LIWC anger

Non-emotion LIWC features have established utility for predicting depression and anxiety. These features capture aspects of symptomatology outside emotion, such as increased self-focus, social isolation, and usage of health-related words. Nonemotion LIWC features would therefore be expected to be complementary to emotion features, and our work confirms that and leveraging both may achieve the best results. We trained a machine learning model using these features in conjunction with emotion features to predict depression (AUROC 0.671) and anxiety (AUROC 0.657). That these models show similar performance using the same features to predict different outcomes may be explained by the large overlap in symptoms between anxiety and depression, e.g. both are characterized by negative self-talk and hopelessness. Additionally, depression and anxiety are often comorbid; indeed, in this dataset, 74.5% of assessments with a GAD-7 score above the diagnosis threshold also had a positive depression finding, and 70.6% of positive anxiety questionnaires also had a positive anxiety finding.

There are important ethical considerations when analyzing patient-generated natural language to infer mental state. Any passive monitoring of patient-generated data may be considered invasive. Due to the sensitive nature of personal health data, such data are subject to protections that do not apply to non-health data. When health-related insights are derived from data that may be neither private nor health-related (e.g. social media posts), obtaining informed consent and handling inferences with appropriate care is paramount. While academic studies such as the current work are governed by rigorous institutional ethics guidelines regarding consent and data sharing, different rules apply to healthcare organizations and commercial

entities. The use of technologies such as the ones presented here may be acceptable if conducted by trusted entities, such as healthcare providers, in order to support care [94]; on the other hand, consumers may be wary of commercial entities conducting such analyses. Further research, as well as applications of the findings presented here, must take such considerations into account.

This work has several limitations. The data used here stem from predominantly female, young, and well-educated participants, and results may therefore not generalize to populations with a different makeup. If predictive algorithms were to be deployed in practice, fairness may be a concern if predictive performance differs for underrepresented groups. In addition, the GoEmotions dataset used to train the BERT-based models is drawn from Reddit, which has been shown to have a disproportionately high representation of young male users [220]. Though it is encouraging that models trained on these data produce features that correlate well with symptom severity in the current study, the development of annotated datasets drawn from a more diverse population may lead to models that better address linguistic and cultural differences in the ways in which emotions are expressed.

Several features used in the random forest classification model are expected to be highly redundant (e.g. GoEmotions Cowen sadness, GoEmotions Ekman sadness, and LIWC sadness; calculated negative emotion variables that are calculated using sadness). However, interdependent features should not affect the random forest's ability to leverage all features optimally to optimize predictive performance.

This work enables and informs future work. We showed that BERT-based emotion features are associated with depression and anxiety status; however, this work did not assess longitudinally if changes in emotion track with changes in depression and anxiety. While existing work demonstrated this relationship for depression-related LIWC features [135], future work may aim to ascertain whether changes in emotion features over time also predict longitudinal patient trajectories. This work also informs feature selection for future work in

depression and anxiety prediction. Emotion variables can be obtained with a range of extraction approaches. Our results indicate the GoEmotions variables may be a better choice than LIWC for emotions. Nevertheless, LIWC features have a place in future work. LIWC's syntactic and topic features were shown in prior work to be associated with depression scores as well as longitudinal patient trajectories and continued to demonstrate utility in this work.

We determined that fine-grained emotions measured in the language of individuals are associated with and predict anxiety and depression status. The associations we found reflect previous findings. This work thus contributes evidence of the reliability of such measurement approaches, supporting the use of these methods in future work investigating the nature of depression and anxiety. For example, these features could aid investigations into depression phenotypes through cluster analysis, as well as psychology research investigating the differential expression of similarly-valenced emotions in depression and anxiety, e.g. by aiding data collection.

Additionally, this work has important clinical implications. Measurement-based care is facilitated by periodic progress assessments, but additional data collection incurs additional workload. In text-based therapy, depression and anxiety status may instead be automatically determined from already-available patient messages. In clinical settings, interpretability is essential; thus, models based on interpretable features such as emotions may be preferred over black-box models classifying raw text directly. Future work may therefore investigate opportunities to leverage emotion-based predictive models for clinical decision support.

5.5 Conclusion

Extraction methods differ in the quality of emotion features extracted. With the data and approaches presented here, emotion features extracted by the GoEmotions BERT-based model not only explained more variance in univariate mixed-effect regressions, but also contributed significantly to predictions of depression and anxiety status by a random forest classifier.

Further, while non-emotion variables obtained from LIWC remain valuable in linguistic modeling tasks, GoEmotions' level of granularity offers clinically relevant nuance that prevailing tools cannot capture.

Chapter 6. From benchmark to bedside: Transfer learning from social media for suicide risk prediction with patient-generated text

Aim 1 (Chapter 3) revealed two opportunities for AI in the context of Caring Contacts. The first opportunity concerns cognitive support for message authoring, such as extracting clinically actionable insights from messages. I developed and published an approach to extracting clinical constructs of depression, which are actionable because they are also therapeutic targets (Chapter 4). I also investigated extracting emotions from patient-generated text, which might inform how clinicians respond to patients (Chapter 5). But Aim 1 also confirmed that suicide risk scoring is promising, as long as it can be used to augment human cognition rather than automate it. In the work described in this chapter, I therefore developed a prioritization model for suicide risk assessment. Rather than automating the clinical decision of whether a given patient needs follow-up, this model is intended to triage incoming messages so urgent ones can be prioritized. I envision a workflow where clinicians still review and respond to all messages, with the ultimate decision of whether to follow-up and how remaining with the human expert; however, the model is leveraged to pre-sort the message queue, so that messages that are more likely to require urgent intervention are reviewed first. I took a utility-oriented approach to establish whether the model could be expected to have practical impact: I developed a utility metric incorporating workflow factors and used it to evaluate my model in terms that are meaningful to stakeholders and can inform decision making regarding model implementation.

Abstract. Suicide is a leading cause of death in the US. Automatic risk detection approaches can facilitate timely intervention. Natural language processing has demonstrated

utility for predicting suicide risk from patient language, with research efforts in this area stimulated by publicly-available performance benchmark sets using expert-annotated social media posts. However, research utilizing data from clinical settings is relatively scarce. Though annotated social media data are relatively abundant, it is not apparent whether these datasets offer utility for clinical settings, given differences in the nature of the communication concerned. Neural transfer learning, through which information learned from training for one task is brought to bear when training for another, is a promising methodology to extend the utility of previously annotated data. However, it remains unclear whether this approach can bridge the gap between social media and clinically derived patient-generated text for automated suicide risk assessment. Additionally, the metrics typically used to evaluate prior approaches do not easily translate to the operational value that may be expected from a model in a real-world deployment. In a clinical dataset, we demonstrate a BERT-based suicide risk prediction approach that uses transfer learning with social media posts to alleviate the limitations of the clinical dataset's small size and demonstrate its anticipated utility using a novel metric. Results show that the approach outperforms baseline approaches and approximates human performance for this task.

6.1 Introduction

Mental health is a critical health issue in our era, and suicide is now a leading cause of death in the United States [30–32]. Caring Contacts is a promising intervention that reduces suicidal thoughts and behaviors, suicide attempts, and suicide completion [44]. However, resource shortages have precluded the broad adoption of this potentially labor-intensive intervention. In intervention designs with two-way communication, patients may reply to messages if they wish, and will receive further tailored support from Caring Contacts staff in response, increasing the level of support but adding workload for care teams.

In Chapter 3, we conducted extensive interviews with clinical and administrative stakeholders experienced with Caring Contacts interventions, revealing opportunities for machine learning methods to alleviate this workload burden and therefore make it possible to provide scalable suicide crisis support; for example, natural language messages sent by patients to intervention staff may be automatically triaged for urgent follow-up [221]. However, the development of predictive models is hampered by the size and availability of clinical suicide risk datasets. Training state-of-the-art neural network-based natural language processing (NLP) models from scratch requires considerable amounts of data [222], and while this has been ameliorated to some degree in recent years by neural transfer learning approaches leveraging unlabeled datasets [223], models perform best with large amounts of training data, i.e. thousands of labeled examples or more. Obtaining this amount of labeled data is typically not feasible in clinical settings, where data privacy and security are carefully controlled [157,158], making it difficult to share labeled datasets for reuse or to collaborate on benchmark dataset development. Resource limitations in healthcare settings may also impede annotation efforts locally. Additionally, while existing real-world healthcare data, such as from electronic health records (EHRs), can be harvested for secondary use, patient-generated natural language data is not commonly collected as part of routine healthcare operations, so those data must be purposely collected, further limiting the size of related datasets.

In contrast, vast amounts of publicly available social media data are continuously generated; they are also more easily annotated because the comparatively lower level of privacy concern allows outsourcing, e.g. via crowdsourcing platforms, or by teams of collaborating expert annotators. Outside of structured interventions, suicidal individuals do reach out to peers in informal settings such as on social media platforms, and a substantial amount of work has now been published on predicting suicide risk from the natural language occurring in social media posts. For example, Coppersmith et al. predicted expert suicide risk from a range of social media data [93], and Shing et al. [126] created a publicly available annotated dataset of Reddit

posts and demonstrated the ability to assess suicide risk in this benchmark dataset. As part of the 2019 CLPsych Shared Task challenge, several machine learning approaches were benchmarked against this dataset, and deep learning techniques demonstrated the most promising performance [125]. The clinical and social media settings differ in many respects, but there is also substantial overlap; for example, message length and formality may differ, but both are natural language written by patients, conveying patients' perspectives in their own words. If datasets, trained models and findings from work conducted with social media could be effectively transferred to clinical risk prediction algorithms, this would contribute substantially to alleviating the data challenges for clinical risk prediction models.

Prior work in the realm of transfer learning suggests that information is transferable between domains and tasks, i.e. it is possible to use data from a different domain or task to augment learning for a task with data scarcity by leveraging available datasets that are related but distinct, e.g. with a different input feature space or data distribution [224]. Selecting a domain-specific pre-trained language model is a well-established approach to leveraging multiple data sources: By choosing a pre-trained model that has been exposed to mental health-related language and screening tasks as a starting point, it is possible to boost performance in related tasks. After pre-training on general-purpose text, such models are further pre-trained on domain-specific corpora. Like the original training of these models, this training is unsupervised (or "semi-supervised", with the language itself providing implicit labels): models are trained to predict held-out ("masked") words, amongst other objectives that do not require the assignment of labels by annotators. For example, BioBERT, a base BERT model further pre-trained on biomedical research literature, improves performance on biomedical tasks compared to base BERT [109]. However, domain-adapted pre-trained models must be further optimized for individual prediction tasks via fine-tuning of model parameters for a particular ML task using labeled data. Multi-stage pre-training and fine-tuning protocols have been proposed, notably by Howard and Ruder [225]. Here, we apply such an approach to harness the power of large,

publicly available social media datasets to mitigate the problem of small dataset size in clinical suicide risk prediction. We hypothesized that neural transfer learning from social media datasets would provide an effective mechanism through which to leverage publicly available social media data, annotated for a benchmark task, for suicide risk prediction in a clinically derived data set.

While artificial intelligence (AI) has the potential to fundamentally transform the way healthcare is delivered, many promising use cases for AI have failed to impact patient care meaningfully. A lack of demonstrated operational utility of existing models has been identified as one of the most significant barriers to achieving the adoption of AI into clinical practice [14]. Developing ways to accurately capture a model's potential clinical utility is therefore essential. Clinical trials are critical to establishing the safety and efficacy of drugs, interventions, medical devices, and more, and represent the gold standard of evidence in biomedicine, including for biomedical AI [226]. Yet, they are underutilized for this purpose, partly because the desire to accelerate adoption is difficult to align with the substantial time and resource requirements of clinical trials [227]. Accordingly, model utility should be established before proceeding with clinical trials. In contrast, standardized machine learning performance metrics, such as the area under the receiver operator characteristic curve (AUROC), are easy to assess, widely used, and generally recommended in machine learning research. While such metrics are indispensable to making model performance comparable across settings, the ultimate value of a predictive model intended for clinical use is determined by its real-world impact on healthcare delivery, which depends on many factors; predictive accuracy is only one of these factors. Evaluation approaches that accurately estimate anticipated model benefits must therefore consider these factors by incorporating those parameters of the deployment environment that can be assessed at development time. For example, Jung et al. assessed the clinical utility of a model recommending advanced care planning, and found that the model's utility was constrained by the health system's capacity to provide this service [15]. Their evaluation helped them devise a

strategy for harmonizing the model with practical constraints to realize optimal utility. Bayati et al. performed a cost-effectiveness analysis of the impact of intervening on the basis of model predictions, incorporating both the cost of a post-discharge intervention and its potential to reduce near-term readmission [22]. The authors estimated that using the system would reduce readmissions by 18.2%, and costs by 3.8%. Here, we develop a utility metric, average time to response in urgent messages (ATRIUM), for a suicide risk prediction model within the Caring Contacts workflow, and use it to evaluate our models.

6.2 Methods

6.2.1 Data

To explore the effect of transfer learning with social media data for the purpose of suicide risk prediction in a clinical dataset, we used the University of Maryland Reddit Suicidality Dataset (Version 2), a dataset of Reddit posts from suicidal individuals curated by Shing and colleagues [125,126], henceforth the social media (SM) dataset. Specifically, we used the portion of the dataset annotated for the task of flagging high-risk individuals amongst Redditors with posts in the /r/SuicideWatch subreddit. Using their posts, user risk levels were rated by crowd workers as well as by 4 experts: a suicide prevention coordinator at the Veteran's Administration, a committee co-chair at National Suicide Prevention Lifelines Standards, a doctoral student with training in suicide assessment and treatment, and an ED psychiatrist. Posts were rated on the following scale: a – no risk, b – low risk, c – moderate risk, d – severe risk. Note that the user-level rating results in all posts belonging to the same user receiving the same class label. The "flagging" task consists of distinguishing unconcerning utterances (a) from those indicating an elevated risk that may warrant intervention (b, c, d). The dataset contains 1105 posts with 222 words each on average, with 82.3% of posts expressing some level of risk. While Reddit does not collect the demographic information of its users, a 2016 report estimated

that 69% of users were male and 58% were 18-29 years old; 63% were white, 10% were black, and 14% were Hispanic [228]; however, these numbers are likely to differ between subreddits.

Our clinical dataset consists of text messages collected as part of the Caring Contacts Via Text (CCVT) clinical trial, which investigated the effectiveness of Caring Contacts to reduce suicidal thoughts and behaviors [40]. In this trial, Comtois and colleagues randomized 658 military service members across three military installations to either standard care, or a text message-based Caring Contacts intervention combined with standard care. All participants reported suicidal ideation at baseline, and 44.3% had previously attempted suicide. Participants in the Caring Contacts condition received 11 text messages over the course of the 12-month intervention period.

In CCVT, patients either do not respond, or respond to scheduled messages with one of the three levels described above. Study staff responded based on their clinical judgment of whether the patient needed urgent support and availability. A patient may respond to an automatically scheduled message, e.g. “Hi there, hope your week has been good!”, with a message indicating distress, for example, “Thanks, but this week has been horrible.” While the study protocol did not prescribe any particular response time frame, study staff treated such messages more urgently than others. If the message indicated an acute crisis, staff immediately followed up with a text message and a phone call, rather than a text message alone. In contrast, a patient may respond positively or neutrally, e.g. “Thank you!”. Most messages of this kind were still responded to, but not with urgency, i.e. with some delay. Staff were notified of every incoming message, and judged urgency as messages were received. After the study concluded, study staff annotated messages for the level of distress expressed by the message author (i.e. message urgency) on the following scale: 0 – none; 1 – difficulty; 2 – non-urgent distress; 3 – urgent distress/crisis. The dataset consists of 1229 messages from 221 unique users, who were 82% male and 25.2 years old on average (standard deviation 6.3). 66.0% were white, 10.0% were African American, 9.1% were Hispanic or Latino, 3.0% were Asian/Pacific Islander, 1.2% were

American Indian/Alaska Native, and 9.1% were Mixed or other. See Comtois et al. [40] for further details on the participant population and study protocol. 18.8% of messages expressed at least some difficulty. Each message contains 9.3 words on average. The two datasets were aligned for the single binary classification task of flagging messages for follow-up, i.e. distinguishing messages labeled "a" from those labeled "b", "c", or "d", and messages labeled "0" from those labeled "1", "2", or "3", respectively. Grouping the classes this way was necessary because of the relative rarity of messages labeled "2" (3.3%) and "3" (1.1%) in the CCVT dataset. This aligns with the intended purpose of the SM "flagging" data subset. Clinician experts confirmed that performing this simplified version of the task is still clinically useful.

Table 6.1. Data characteristics. *Pew Research Center estimates

	SM	**CCVT**
Context		
Setting	Social media platform Reddit	Clinical trial among military service members
Population	Reddit users proactively seeking advice online	Individuals identified by clinical experts to be at suicide risk
Message audience/confidentiality	Anonymous post to the general public	Initiated in a confidential one-on-one setting with a suicide prevention professional; text messages from/to personal mobile phones
Purpose of messages	Social media posts e.g. to seek advice or empathy	Caring Contacts
Participant characteristics		
Number of participants	621	221
Sex male	69%*	82%
Age	58% 18-29 * 33% 30-49 * 7% 50-64 * 1% 65+ *	Mean 25.6 (SD 6.3)
Race/Ethnicity		
White	63%*	66.0%
Black	10%*	10.0%
Hispanic	14%*	9.1%
Message characteristics		
Message count	1105	1229
Indicating risk/urgency	82.3%	18.2%
Words in message, mean (SD)	222.0 (250.9)	9.3 (10.2)

Distribution and data characteristics differ between these data sets (Table 6.1). Texts were produced by different people (clinically identified individuals at suicide risk in a population of military service members vs. Reddit users proactively seeking advice online), in a different setting (confidential one-on-one setting with a suicide prevention professional within a clinical trial vs. an anonymized suicide watch forum available to all members of the general public), for a different purpose (Caring Contacts vs. a range of social media motivations such as seeking advice or empathy). The documents are significantly longer (mean length of 222 vs. 9 words) and the class distribution differs significantly (82.3% vs. 18.8% positive, i.e. 5:1 vs 1:5) for SM versus CCVT respectively. These differences pose challenges for transfer learning. Serendipitously, the two datasets also have important similarities that transfer learning may be able to capitalize on: both contain text written by an individual pre-disposed to suicidal thoughts and behaviors, population demographics are similar, and even though label definitions have different nuances, the labels can be grouped in the same way (i.e. no risk vs. any level of risk).

6.2.2 BERT model

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning architecture for NLP reported by Devlin and colleagues at Google in 2019 [100]. Using natural language documents as inputs, the BERT model develops representations for words based on the contexts in which they appear; notably, in contrast to prior models, both the preceding and subsequent text is used (bidirectionality). BERT was initially trained on large amounts of natural language in an unsupervised (or "semi-supervised") manner, learning to predict held out ("masked") words and the sequence in which sentences occur. This pre-training informs the contextual representations it derives from previously unseen text, providing both initial representations of words (or their components) and initial weights through which to estimate their influence in context. These contextual representations can then be used toward a wide

range of prediction tasks, including text classification, part-of-speech tagging, and question answering.

When using BERT for new tasks, it can help to use a model pre-trained on a corpus with domain relevance. Several domain-specific pre-trained BERT models are available for use; for example, BioBERT, trained on biomedical research literature, improves upon base BERT in biomedical tasks [109]. Here, we used Public Health Surveillance (PHS)-BERT [229] as a base model because it has a demonstrated performance advantage on suicide-related prediction tasks compared to other biomedically relevant pre-trained models such as BioBERT and the mental health-specific MentalBERT [111]. Additionally, in preliminary exploratory work, PHS-BERT outperformed other pre-trained BERT models in our setting.

6.2.3 Utility metric: Average Time to Response In Urgent Messages (ATRIUM)

We aimed to develop a way to measure clinical utility in terms of average time to response in urgent messages (ATRIUM). The metric assumes that the clinical setting is such that we have work capacity to respond to a set number of messages quickly, i.e. prioritize them, and respond to the others at a later time, i.e. deprioritize them. The algorithmic problem is to assign messages to the prioritize bucket and the deprioritize bucket, such that as many urgent messages as possible get prioritized, and as few urgent messages as possible get deprioritized. Thus, a machine learning model aiming to assist in the message triage process should assign each message a score such that urgent messages get a higher score than non-urgent messages; we then take the top ranked messages and prioritize them. Assuming a set work capacity for prioritization, i.e. a capacity of k, the model's accuracy might be measured in terms of precision at k. However, we are also interested in minimizing the number of urgent messages that would incorrectly be deprioritized, i.e. the number of urgent messages that did not make it into the top k predictions. We therefore define two parameters: a, the number of urgent messages within the top k; and U, the total number of urgent messages. Note that a/k is equivalent to the precision

at k, and the number of missed urgent messages is $U - a$. U depends on the data and is therefore treated as a constant here; a is specific to the model. k, in practice, is a tradeoff between costs and benefits, i.e. the implementing healthcare organization's staff availability compared to the amount of time that can be saved, and will depend on the organization.

Let T_{urgent} and $T_{non-urgent}$ be the response times that can be expected for prioritized and de-prioritized messages, respectively. We define ATRIUM to be the average response time across urgent messages, given that a of them will correctly be treated with urgency, i.e. responded to with a delay of T_{urgent}, and the remaining U-a messages will incorrectly be treated without urgency, i.e. responded to with a delay of $T_{non-urgent}$. In other words, ATRIUM is a weighted average of quick and slow response times, as follows:

$$ATRIUM(a, U) = \frac{aT_{urgent} + (U - a)T_{non-urgent}}{U} = \left(\frac{a}{U}\right)T_{urgent} + \left(1 - \frac{a}{U}\right)T_{non-urgent}$$

We drew upon the observed response times in the CCVT clinical study dataset for messages annotated as urgent and non-urgent to define the random variables T_{urgent} and $T_{non-urgent}$. In other words, we assume that in our dataset, all urgent messages were treated with priority, and all non-urgent messages were deprioritized; in other words, $k = a = U$. We use the observed response times for messages annotated as urgent as the basis for T_{urgent}, and the observed response times for messages annotated as non-urgent as the basis for $T_{non-urgent}$. These response times therefore capture the human baseline for response times, i.e. what might be expected in a setting where all urgent messages are responded to quickly, and all non-urgent messages are responded to slowly.

We use the set of observed response times to empirically define the probability distributions for T_{urgent} and $T_{non-urgent}$, respectively. We use all messages where paired timestamps were available, i.e. where a patient replied and staff responded to the reply with another text message. Some responses were sent after more than 600 minutes; we determined via manual review that reasons included technical errors (message did not go through) and

responsible staff being out of the office for the day (response sent between 8am and 9am the following day). For this analysis, we consider these to be non-representative of the response time distribution and therefore exclude them. 421 patient messages with timestamped responses remained, with 79 (18.8%) expressing some level of difficulty, and 321 (81.2%) being positive or neutral. Response times were approximately normally distributed for each class, with a slight skew towards earlier times for urgent messages. The average response times were $\overline{T_{urgent}} = 106.2\ min$ and $\overline{T_{non-urgent}} = 130.6\ min$, respectively.

6.2.3.1 Random baseline

If there were no triage, and response times were instead assigned randomly, the number of correctly predicted urgent messages among the top k predictions, a, would follow a hypergeometric probability distribution, which describes the probability of a successes in k draws, without replacement, from a finite population of size N that contains exactly U objects of interest, and has the following expected value:

$$a = k\left(\frac{U}{N}\right)$$

At each k, we therefore consider this to be the value of a that would be seen in a random ranking of messages (i.e. no triage), to be improved upon by any alternative approach.

6.2.3.2 Human baseline

As mentioned above, the theoretical maximum capacity for the number of messages that can be treated with priority depends on staff availability. The CCVT trial was staffed such that staff shortages would not be responsible for patients in urgent need receiving a delayed response. We therefore assume that, in this dataset, every urgent message was recognized as urgent and handled accordingly (recognizing that response times will vary due to many real-life factors within the empirical probability distribution), and every non-urgent message was recognized as such and handled without urgency. We thus assume that the capacity matched the

number of urgent messages in the dataset. In this dataset, there were 79 urgent messages, i.e. $k = a = U = 79$. ATRIUM is defined such that it is equivalent to the observed data at $k = 79$ (see Table 6.3). For different values of k, we define a for the human baseline as follows:

$$a = \begin{cases} U \; if \; k > U \\ k \; otherwise \end{cases}$$

6.2.3.3 Calculation of expected values for ATRIUM

To calculate the expected value of $ATRIUM$ for each model, we first determine a at each $k \in \{1, \dots, N\}$ where N is the total number of messages in the dataset. This is accomplished as described above for the human and random baselines; for the machine learning models, we rank order the set of 421 predictions corresponding to timestamped messages, and count the number of correct instances in the top-ranked k predictions, for each k. Next, we determine the expected values for ATRIUM for each model, given U and a for each k, using expected urgent and non-urgent response times from previous or expert knowledge. We use $T_{urgent} = 106.2 \; min$ and $T_{non-urgent} = 130.6 \; min$ as point estimates.

6.2.4 Model development

We trained and evaluated a bag-of-words baseline model, a BERT-based suicide risk classifier, and a BERT-based suicide risk prediction classifier that additionally leverages supervised transfer learning from an annotated social media dataset, using cross-validation.

6.2.4.1 Training and performance calculation using cross-validation

Models were trained and predictions for all examples were obtained using 5-fold cross-validation. Data were split into 5 subsamples in a stratified manner by user, such that all documents from a single participant appeared in the same fold and the label frequency was approximately retained within each fold. In each of 5 rounds, 4 folds of the data were pooled and used to train the model as described below, and predictions are produced for the remaining fold.

We train 5 models in this way to obtain one prediction for each instance in the entire dataset. Reported performance metrics are calculated on this set of predictions. We ranked only the predictions for messages for which paired timestamps were available in order to calculate a, and calculate ATRIUM, i.e. the average time to response across all urgent messages.

This process was repeated 5 times, randomly splitting the data into 5 folds differently each time. Reported metrics refer to the median performance across these 5 different randomizations.

6.2.4.2 Bag of words baseline

We trained and evaluated a bag of words baseline model following a standard approach that has been widely used in clinical applications on account of its ready interpretability. We removed stop words, punctuation, and special characters from each document and lemmatized words using Natural Language Tool Kit (NLTK) [230]. Then, for each document j, we construct a feature vector $b_j = [w_0, w_1, \ldots w_n]$ where n is the number of unique words in the corpus and w_i is the count of that word in document j. We then fit a logistic regression model, support vector machine, and random forest model and determine performance across the entire set using the cross-validation data splits and training procedure described in the preceding section. Hyperparameters were selected using Scikit-Learn's GridSearchCV [210], applied to a validation partition (10%) of each training split.

6.2.4.3 BERT model

The first set of BERT models was trained by fine-tuning PHS-BERT (from the Huggingface library of pre-trained models [231]) on the CCVT dataset using the 5-fold cross-validation procedure described above. We trained on 90% of each training split for up to 12 epochs using a learning rate of 10^{-6}, and select the best-performing epoch according to the cross-entropy loss calculated on the remaining 10% of the training data (the validation set). To

account for class imbalance, the loss function (cross-entropy) was modified to weight classes according to the reciprocal of their frequencies in the training data, i.e. approximately 1:5.

In accordance with common practice, we initially selected a small learning rate of 10^{-5}. The learning rate of 10^{-6} was selected after we observed that this initial learning rate resulted in the cross-entropy loss calculated on the validation set increasing after only a single epoch of training. This selection of a lower learning rate is consistent with the common practice of reducing the learning rate in neural transfer learning to mitigate the loss of knowledge from prior pre-training due to overly aggressive fine-tuning, also known as catastrophic forgetting.

6.2.4.4 BERT model with transfer learning

The second BERT model was trained in two stages. We first fine-tuned a PHS-BERT model with 90% of the SM dataset (training set) for up to 12 epochs using a learning rate of 10^{-5}, carrying forward the best-performing epoch as determined using the remaining 10% of the data (validation set). We then followed the 5-fold cross-validation procedure described above to continue the training process using the CCVT training set, training for up to 12 epochs using a learning rate of 10^{-5}, again selecting the best-performing epoch according to the performance metrics achieved on the validation set as before. To account for class imbalance, the loss function was modified to weight classes in both phases according to the reciprocal of their frequencies in the respective training datasets, i.e. approximately 5:1 and 1:5 for the SM and CCVT datasets, respectively.

6.2.4.5 BERT model with transfer learning using Howard & Ruder learning rates

We additionally trained a BERT model using a contemporary approach to selecting learning rates to optimize multi-stage transfer learning, described by Howard and Ruder [225]. This involves three techniques – discriminative fine-tuning, slanted triangular learning rates (STLR), and gradual unfreezing – aimed at counteracting catastrophic forgetting. Discriminative fine-tuning involves selecting learning rates that are exponentially larger for later layers than

earlier ones. Tuning each layer with a different learning rate in this way allows more prior knowledge to be retained in earlier layers. STLR involves linearly increasing the learning rate for a set number of training iterations, and then linearly decreasing it for the remainder of the training iterations, allowing the model to first select a general region of the parameter space, and to then slowly converge within that region. These two techniques are combined for the first transfer learning phase. In the second transfer learning phase, i.e. the classifier fine-tuning phase, we additionally employ gradual unfreezing. Here, we first freeze the layers of the model, i.e. set the learning rates to zero, and then unfreeze one layer per epoch, starting with the last layer. For a detailed description of these techniques, see Howard and Ruder [225].

In accordance with the multi-phase fine-tuning approach by described by Howard & Ruder, we used the first two techniques for fine-tuning the PHS-BERT model using the SM dataset in the first transfer learning phase, using 12 epochs and a target (maximum) learning rate for the last layer of 10^{-5}. As per Howard & Ruder's recommendation for the values of the scheduling parameters, each layer's target rate is the previous layer's rate multiplied by 2.6; within each layer, the learning rate is linearly increased starting from $1/32^{th}$ of the target rate, until it reaches the target rate after 10% of iterations, and is then linearly decreased until it reaches the starting point again at the last iteration. We combined all three techniques in the next transfer learning phase to further tune the model to the CCVT data set, using 12 epochs and a target learning rate of 10^{-3}; that is, all layers except the last one were frozen in the first epoch, all layers except the last two were frozen in the second epoch, and so on. This rate was selected after the model did not seem to improve from epoch to epoch with smaller rates, as evidenced by the cross-entropy loss on the validation set. A requirement for larger learning rates should perhaps be expected in this approach, which further constrains the amount that can be learned by partially freezing the model. We additionally incorporated class weighting as before.

6.3 Results

Our experiments were designed to answer two main questions. The first was whether transfer learning from the social media set would improve performance with our clinically-derived data. As shown in Table 6.2, all models achieve competitive AUROCs. Precision and recall are well balanced in all neural models, possibly a result of class weighting. In terms of F1 score, the best transfer learning approach achieved a 6.3% (0.063) improvement over the BERT model that did not use transfer learning, and a 21.2% (0.212) improvement over the best bag-of-words baseline model. The best transfer learning approach, which used learning rate scheduling optimized for transfer learning proposed by Howard and Ruder, improved upon the baseline transfer learning model by 1.3% (0.013). These are substantial improvements in performance for deep learning models over classical machine learning approaches, and for the BERT models with transfer learning over their counterpart trained without this step, and for the model with customized learning rate schedules over all others.

Table 6.2. Median performance metrics across 5 runs with different cross-validation splits, calculated on aggregated predictions on the test splits.

Precision@79 represents an idealized scenario in which staff are available to triage exactly the number of urgent messages in the set.

	Acc.	F1	Pr.	Rc.	AUROC	AUPRC	Pr. @79
Bag of Words							
Logistic regression	0.868	0.585	0.687	0.509	0.858	0.660	0.595
SVM	0.854	0.581	0.607	0.558	0.801	0.542	0.595
Random Forest	0.858	0.455	0.753	0.326	0.890	0.656	0.608
PHS-BERT + CCVT	0.902	0.734	0.721	0.741	0.947	0.825	0.722
PHS-BERT + SM + CCVT	0.924	0.784	0.798	0.768	0.955	0.860	0.785
PHS-BERT + SM + CCVT with Howard & Ruder LRs	0.925	0.797	0.793	0.772	0.961	0.875	0.785

The second question concerned the extent to which improvements in classification performance would be reflected by a measure of clinical utility. In Table 6.3 and Figure 6.1, we present the estimates for the time saved on average per urgent message due to the use of each model as a better (i.e. more relevant to a clinical setting) approximation of model performance.

The difference between ATRIUM in a scenario without triage (random baseline) and the ATRIUM achievable with predictive model use is shown in parentheses in Table 6.3.

Table 6.3. Average time to response in urgent messages (ATRIUM) (time saved compared to baseline), calculated using the model with the median performance metrics.

k of 79 represents an idealized scenario in which staff are available to triage exactly the number of urgent messages in the set.

k	20	50	79	100	120	150
0. Random baseline	129.6 (0.0)	127.8 (0.0)	126.1 (0.0)	125.0 (0.0)	123.8 (0.0)	121.7 (0.0)
1. Bag of words (LR)	125.1 (4.5)	119.3 (8.5)	116.2 (10.0)	114.6 (10.4)	112.4 (11.4)	111.6 (10.1)
2. PHS-BERT + CCVT	124.4 (5.1)	117.0 (10.7)	113.1 (13.0)	111.0 (14.1)	109.1 (14.8)	108.2 (13.5)
3. PHS-BERT + SM + CCVT	124.4 (5.2)	116.8 (11.0)	111.6 (14.5)	109.5 (15.6)	107.5 (16.4)	106.9 (14.8)
4. PHS-BERT + SM + CCVT with Howard & Ruder LR	124.3 (5.3)	116.7 (11.1)	111.5 (14.6)	109.1 (15.9)	107.6 (16.2)	107.5 (14.2)
5. Human triage	124.4 (5.2)	115.4 (12.4)	106.2 (19.9)	106.1 (18.9)	106.2 (17.6)	106.2 (15.4)

The random baseline model represents the time to response that would be expected if no triage was done, i.e. whether a message was treated with urgency was random. The bag of words model markedly improves upon this random baseline, saving approximately 10.4 minutes for each urgent message with $k = 100$. The non-transfer BERT model outperforms the bag of words baseline, saving an estimated 14.1 minutes per urgent message. The best performing model is the one leveraging transfer learning with Howard and Ruder's strategy for selecting learning rates, which saves about 15.9 minutes per urgent message over the no-triage baseline at $k = 100$. Response times for urgent messages that might be expected when using this classifier to perform automatic triage closely resemble what might be expected if a human being were to conduct ongoing triage: approximately 18.9 minutes saved per urgent message over random triage at $k = 100$. These estimates show that the advantages in classification performance shown in Table 6.2 are indicative of faster response times, of a magnitude that suggests the potential for real clinical impact.

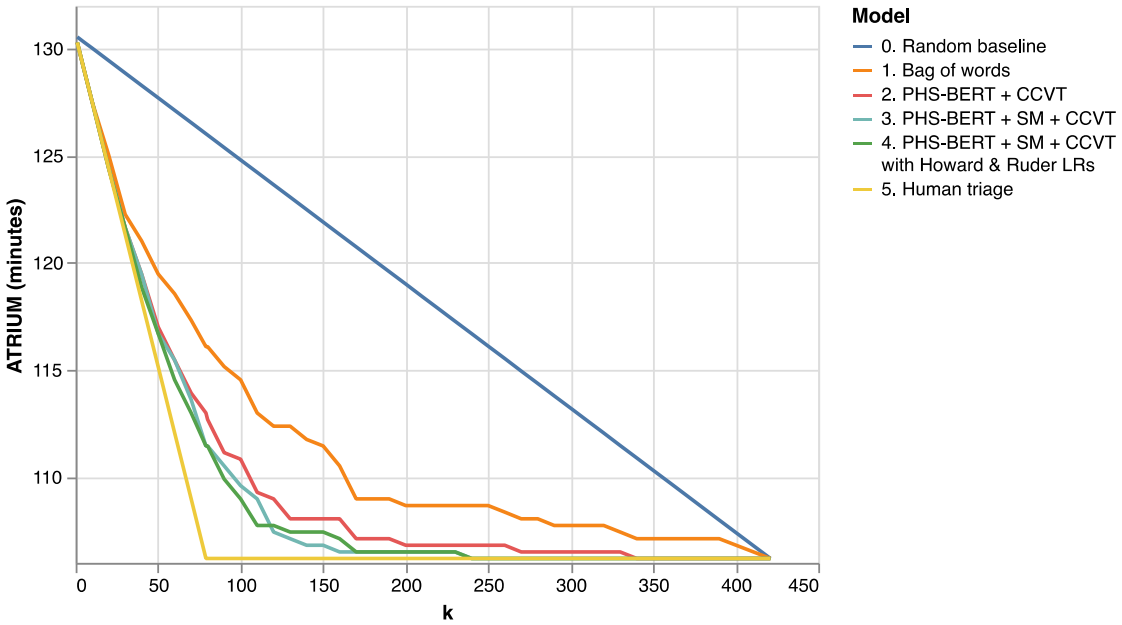

Figure 6.1 Average time to response in urgent messages (ATRIUM) vs. k.
"Bag of words" is the best bag of words model by F1 score, which used Logistic Regression.

As expected, more time can be saved if k is set higher; this is true even for the random triage model. The time saved over random triage therefore begins to decline at the (artificial) optimum with $k = 79$ for the human triage model. The machine learning models do not have this artificial peak so this pattern does not hold for them; this is an artifact of the data used in our approach to creating the human baseline, which has 79 positive examples.

6.4 Discussion

This work found that transfer learning with annotated social media data developed to serve as a benchmark for a shared task evaluation to promote research improves suicide risk classification performance with messages collected in the context of a clinical intervention. We also developed an evaluation approach that estimates clinical utility in a way that is intuitively more meaningful to clinical stakeholders than conventional metrics.

Our findings demonstrate that transfer learning has a significant impact on performance. In our case, we not only use a BERT model augmented by additional pre-training

with demonstrated success for suicide risk prediction [229] from the Huggingface library of pre-trained models [231], but also leverage a dataset of labeled social media posts for additional transfer learning. Though there are similarities, the distribution and characteristics of the SM data differ from the clinical data that is the target of our final model. We found that incorporating this social media data markedly boosts the F1 score, from 0.734 to 0.797. This finding demonstrates that combining small clinical datasets with more easily available non-clinical datasets can benefit clinical prediction tasks, despite differences in data characteristics and distribution.

We also developed a utility metric for models employed for message triage within the Caring Contacts workflow. When evaluating the utility of models, it is important to assess not only standardized performance metrics (AUROC, AUPRC), but also the performance metrics that are most relevant for the setting of the problem (Precision at k). To move further along the continuum from "in-situ" model performance to clinical utility, also it is important to assess the model's practical value in terms of quantities relevant and intuitively meaningful to stakeholders. Though impressive, it is not obvious what practical difference the improvement in F1 score would make for the model user. This is in part because the F1 score measures model performance in terms of class predictions (rather than probabilities as a balanced tradeoff between positive predictive value and sensitivity). However, allowing the dynamic selection of a classification threshold is more appropriate in a clinical setting where sensitivity may be valued over specificity or vice versa. Assessing precision at k for a range of values for k is more insightful, as it allows stakeholders to consider the tradeoff between model performance at k and the cost of achieving that particular k (work capacity which depends on staffing), aiding in decision making for implementation design. Yet, precision at k may not be easily understood by clinical stakeholders, and does not capture the magnitude of effect on the workflow that can be expected. Here, we demonstrate one approach to estimating this clinical impact in terms of time saved per urgent message in minutes. We estimated that, compared to employing no triage, on

average approximately 19.9 minutes could be shaved off each urgent message response time if this model were to be employed for automatic triage of incoming messages (at $k = 79$). This is within 1 minute per message of the expected time saved with human triage.

This metric can also inform allocation of staff needed when scaling the intervention in order to achieve a maximum time to response. Scaling allows organizations to benefit more patients, but may incur prohibitive staffing requirements if human beings were responsible for all triage. Assessing ATRIUM for each triage approach across a range of values for the work capacity k allows comparisons of staffing requirements for each triage approach, given an acceptable average response time. For example, if an organization were to determine that an ATRIUM of at most 110 min should be targeted, using the bag of words model would require enough staff to reach a work capacity of almost 170; in contrast, the best BERT model might accomplish an ATRIUM of 110 min at a work capacity of only k=90, corresponding to a 47% reduction in staffing needs.

6.4.1 Ethical and practical considerations for suicide prevention on social media vs. healthcare settings

Our lives are increasingly digitized, with immense amounts of data automatically created or collected by consumer technologies. Smartphone logs, sensor data, and social media data – known as patient-generated data in the health informatics space – have enabled a proliferation of technologies promising to benefit human health within the consumer realm. Social media data, in particular, hold unprecedented promise; they have mediated the development of new markers of mental health symptomatology in depression and anxiety [205,206,208,232], and Facebook have developed and operationalized machine learning to detect users with high suicide risk on their platform [95,96].

These data also represent a unique opportunity to benefit clinical care by improving provider understanding of patient health: recent work, including our own, has demonstrated the

potential of utilizing patient-generated natural language data in clinical settings, showing that language indicators extracted from logs of message-based clinical psychotherapy sessions can predict patient trajectories in depression [137,190].

Indeed, leveraging these data for clinical purposes may be more ethical, practical, and acceptable to users. Barnett and Torous argue that clinicians, not advertising companies, should have the responsibility to identify and help suicidal individuals [98]. Besides the ethical and privacy concerns surrounding publicly traded companies inferring sensitive health information about their users [99,233] and acting on it without explicit consent [93], commercial entities operating such technologies may not be able to respond appropriately to an emergent crisis [98]. Some have warned that bad actors may purposely target individuals identified as vulnerable in public forums [97,233]; others have argued that only medical professionals should engage in triage and intervention [98,99]. Model developers at Facebook humbly state, "Our expertise at building social networking and scalable software systems in no way qualified us to reinvent suicide prevention" [95]. Undoubtedly, trained clinical professionals are best equipped for suicide prevention.

Integrating emergent signals into the clinical workflow to support diagnosis and treatment by medical professionals is a path forward. Because intervention decisions remain with the qualified clinician, such use of patient-generated data is perhaps most responsible and effective. Yet, the intersection between social media research and clinical decision support research remains small. More work investigating how to translate the advances made with consumer technology to clinical settings is therefore needed.

The current research advances our understanding of how best to leverage mental health insights originating outside the healthcare setting as part of ongoing clinical care. We demonstrated an approach to applying the advances made with social media data toward empowering qualified clinicians to deliver better care. Thus, this research helps realize the

translational potential of social media derived signal to detect suicide risk in the context of an established clinical relationship.

6.4.2 Limitations

The CCVT dataset is comparatively small and specific to the Caring Contacts intervention. Nonetheless, it reflects a real clinical use case for suicide risk prediction. It is likely that different sites will have to develop their own models customized to their Caring Contacts participants; however, our approach demonstrates the feasibility of using larger, more easily available datasets to optimize performance on these small ones.

We have made a number of assumptions in order to estimate clinical utility. First, we assume that the CCVT dataset is a "gold standard" representing what human triage can achieve; however, response times were not a primary outcome of the trial and should therefore be considered noisy. While the study protocol included guidance for staff with respect to clinical judgment and response timing, and staff made every effort to follow this guidance, there was no explicit instruction to respond to certain messages more quickly than others – other than messages indicative of an acute crisis, which were addressed by immediately reaching out to participants via phone call. Staff triage and response would have been limited by many practical constraints, e.g. work hours. We did our best to filter the dataset via manual review of response times that seemed non-representative, but this does not fully alleviate the problem. If the original study protocol had explicitly encouraged staff to respond as quickly as possible to messages expressing any level of adversity and deprioritize all others, there would likely be a larger difference between response times for the two groups in this dataset.

Second, the distribution of response times varied significantly even within each class, with substantial overlap between the distributions. In a real clinical practice setting, staff estimates of severity would likely be more fine-grained than "urgent" vs. "not urgent". In this

study, we grouped messages into just two urgency levels, resulting in the loss of some of this granularity and blurring the boundaries between groups.

Third, the choice of summary statistic used to estimate workflow factors may influence utility calculations. For example, in our case, compared to the average response times of 106.2 min and 116.2 min for urgent and non-urgent messages, respectively, the median response times were lower (88.6 min vs. 116.2 min). Here, the use of the median, rather than the average, did not substantially change calculated utilities: the bag-of-words model saved 10.2 min over the random baseline, the best transfer model saved 22.3 min over the random baseline, and the human triage approach saved 22.7 min over the random baseline. However, in practice, the choice of summary statistic may be meaningful and should be carefully considered to avoid over- or under-estimating utility. Finally, the trial was well-staffed to support the number of enrollees, and may represent a best-case scenario for staffing constraints, which are likely to be more severe in settings where funding to support personnel is more limited. Nevertheless, we believe that our aim of proposing a potentially practical and interpretable metric and using it to *estimate* the clinical utility is achieved: even if it is only an estimate subject to several assumptions, it is more interpretable than another performance metric that may mean very little to clinical stakeholders.

6.4.3 Future work

Further optimizations to the presented training process are possible. For example, it is possible that additional hyperparameter tuning, such as an exhaustive search for optimal hyperparameters in the Howard & Ruder learning rate selection strategy (target learning rates, division factor, number of iterations of learning rate growth, etc.), would further improve performance.

The best-performing model presented here would be expected to perform well in the clinical setting that produced the CCVT dataset. We are currently developing an informatics-

supported pilot implementation of Caring Contacts, which includes this model as a clinical decision support component.

6.5 Conclusion

Advances in suicide prediction work using social media data are promising, but the growing divide between cutting edge NLP research and the realities of clinical practice raises questions of the applicability of models emerging from this research to clinical settings. Yet, this is where such models have the highest potential to produce tangible improvements ethically and effectively. We demonstrated the feasibility of offsetting the limitations of the small size of clinical datasets with neural transfer learning using related, more easily available non-clinical data from a publicly-available benchmark; further, we demonstrated practical value to clinical practice using the novel time utility metric. Thus, this work contributes toward bridging the historical implementation gap by translating state-of-the-art NLP advances to data from clinical settings and formally estimating utility toward improved clinical care.

6.6 Acknowledgments

This work was supported by Innovation Grant "Informatics-Supported Authorship for Caring Contacts (ISACC)" from the Garvey Institute for Brain Health Solutions.

This work was in part supported by the Military Suicide Research Consortium, an effort supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-10-2-0181. The views expressed herein are those of the author(s) and do not reflect the official policy or position of the U.S. Army Medical Department, Department of the Army, Department of Defense, the U.S. Government, or the Military Suicide Research Consortium.

The University of Maryland Reddit Suicidality Dataset was provided by Dr. Philip Resnik. We acknowledge the assistance of the American Association of Suicidology in making the dataset available.

Chapter 7. StayHome: A FHIR-native mobile COVID-19 symptom tracker and public health reporting tool

Aim 1 revealed that the workload burden imposed by Caring Contacts combined with a lack of financial resources in preventive care is a barrier to adoption. Additionally, data and workflow integration, e.g. as accomplished by leveraging data standards such as FHIR, are essential for successful digital tools for Caring Contacts.

The work described in this chapter introduces FHIR-native, an approach to FHIR data modeling and FHIR-based application development that maximizes reliance on data standards and tools developed and maintained by the standards community. This reduces the need for customized data modeling and infrastructure development. In this way, FHIR-native enables rapid application development and facilitates cost-effective deployments built on freely available, open-source tools, while ensuring out-of-the-box interoperability. This work was conducted early in the COVID-19 pandemic: the first US case was reported in February 2020, and we began work on this project the same month. Amidst rapidly changing public health guidance and stay-at-home orders, we recognized the urgent need for digital tools to support symptom tracking. By using the FHIR-native approach, we were able to design, develop, and deploy StayHome within just a few months.

The FHIR-native approach is also uniquely suited for suicide prevention applications. Besides meeting the core requirement of data and workflow integration through the use of FHIR, this approach is advantageous for prevention interventions and other settings where minimizing development, maintenance, and deployment costs is of the essence. However, it requires that all data and state required by the application can be represented using FHIR resources. Fortunately, this is the case, as established in Chapter 8.

A version of this chapter was previously published by the Online Journal of Public Health Informatics (OJPHI) as an open-access article. © the authors.

Burkhardt H, Brandt P, Lee J, et al. StayHome: A FHIR-Native Mobile COVID-19 Symptom Tracker and Public Health Reporting Tool. *Online J Public Health Inform*. 2021;13(1). doi:10.5210/ojphi.v13i1.11462

Abstract. As the COVID-19 pandemic continues to unfold and states experience the impacts of reopened economies, it is critical to efficiently manage new outbreaks through widespread testing and monitoring of both new and possible cases. Existing labor-intensive public health workflows may benefit from information collection directly from individuals through patient-reported outcomes (PROs) systems. Our objective was to develop a reusable, mobile-friendly application for collecting PROs and experiences to support COVID-19 symptom self-monitoring and data sharing with appropriate public health agencies, using Fast Healthcare Interoperability Resources (FHIR) for interoperability. We conducted a needs assessment and designed and developed StayHome, a mobile PRO administration tool. FHIR serves as the primary data model and driver of business logic. Keycloak, AWS, Docker, and other technologies were used for deployment. Several FHIR modules were used to create a novel "FHIR-native" application design. By leveraging FHIR to shape not only the interface strategy but also the information architecture of the application, StayHome enables the consistent standards-based representation of data and reduces the barrier to integration with public health information systems. FHIR supported rapid application development by providing a domain-appropriate data model and tooling. FHIR modules and implementation guides were referenced in design and implementation. However, there are gaps in the FHIR specification that must be recognized and addressed appropriately. StayHome is live and accessible to the public at https://stayhome.app. The code and resources required to build and deploy the application are available from https://github.com/uwcirg/stayhome-project.

7.1 Introduction

SARS-CoV-2, a novel coronavirus which causes the disease COVID-19, was first reported in December 2019 in China [234] and quickly spread globally, causing millions to contract the virus and fall ill. In the US, the disease was first discovered in a Washington State nursing home in February 2020. As of August 2020, the virus has claimed over 170 thousand American lives [235]. Starting in April 2020, states imposed severe limitations on businesses and individuals in an attempt to curb spread. As of August 2020, states have relaxed stay-at-home orders and are experiencing diverse outcomes from reopened economies. Blueprints to minimize the increased outbreaks include public health efforts such as widespread testing, identification of cases, and intensive contact tracing [236]. States are rapidly scaling up the number of individuals conducting contact tracing investigations to meet this need, which means that the contact tracing workforce is comparatively inexperienced, yet meets with high caseloads. An additional exacerbating factor is that many sources of essential data, such as community-based testing, medical records, and public health records are managed in ways that inhibit their effective linkage. Informatics solutions may ease and support this labor-intensive, manual work. We can draw on the experiences of patient-centered systems in clinical informatics, which recognize that often the patient and their reported experiences may be the best links between information managed in disparate locations and systems. In the context of COVID-19, this may include events such as testing and clinical visits, as well as personal information such as contacts and travel, and clinically relevant symptom monitoring.

Individuals are encouraged to wash their hands, wear masks, maintain physical distance from others, and self-monitor for symptoms in order to recognize when further steps become necessary. Proactively tracking symptoms may allow recognition of emerging symptomaticity, allowing earlier self-isolation and reducing transmissions, and is therefore a core need of the community. Maintaining a diary of symptoms, possible exposures, virus and antibody testing,

and travel history has the additional potential benefit of providing public health agencies with reliable information to conduct case investigations, should that need arise.

The more people keep and are willing to share such records, the greater the benefit may be to the community. Symptom trackers and exposure diaries should therefore be easy to use and accessible to everyone, including vulnerable populations. Choosing a mobile-accessible design supports this goal due to the ubiquity of mobile phones and the potential to reach more individuals from vulnerable communities. In 2019, over 80% of individuals and over 50% of older adults in the United States owned a smartphone; 17% had internet access only on their phone, with higher proportions for poor populations (26%) as well as Black (23%) and Hispanic (25%) minority groups [237].

Individuals should be able to share diaries of COVID-relevant observations directly and easily with the relevant agencies if they so choose. Using health data standards can lower the barrier to sharing symptom tracker and exposure diary data with third parties. The Fast Healthcare Interoperability Resources (FHIR) [26] standard may facilitate data sharing, both with electronic health records (EHRs) for use in clinical care, and with public health to benefit contact tracing.

7.1.1 Representative COVID-19 apps

The COVID-19 pandemic has prompted the development of numerous consumer applications intended to support individuals, to complement traditional public health surveillance efforts, or both. Some examples of these apps, grouped into broad categories, are shown in Table 7.1. As a longitudinal symptom tracker, StayHome represents a less common application type, in contrast with “low-tech” symptom screeners and “high-tech” contact tracing apps. Yet, there have been examples of such applications seeing high adoption and even enabling disease research: A symptom tracker launched in the UK on March 24, 2020 (“COVID

Symptom Study App") gained 700,000 users within 24 hours and ultimately contributed to the reporting that loss of sense of taste or smell is associated with COVID-19 infection [238–240].

StayHome first and foremost supports self-monitoring by individuals. Supporting public health efforts is a secondary goal enabled by StayHome if a need arises.

7.1.2 FHIR

Lack of standardization can both delay access and reduce the quality of the data available to public health agencies. To address broad issues of health data interoperability, Health Level Seven International (HL7) has published the FHIR standard. Applicable healthcare organizations are required to implement and maintain the FHIR application programming interface (API) per the Interoperability and Information Blocking Rule of the 21st Century Cures Act (effective June 2020) [145,146]. FHIR describes a standard representation for many common entities in the healthcare domain, defines relations between these entities, and describes a variety of computational methods for operating on these entities. The FHIR specification describes recurrent healthcare application business requirements in dedicated modules, accessible via the documentation index on the FHIR website (https://www.hl7.org/fhir/). Two examples are the Workflow Module and the Clinical Reasoning Module. The Workflow Module describes how care plans and specific care-related workflows can be characterized, scheduled, and executed. The Clinical Reasoning Module provides a mechanism to represent and evaluate clinical knowledge in an entirely FHIR-native way, that is, using FHIR-compliant data structures, operations, and logical expression languages like FHIRPath [241] and the Clinical Quality Language (CQL).

Table 7.1 Comparison of COVID app modalities

	Description	Example	Data collected	Longitudinal
Symptom checker / screener	Allows the user to enter their symptoms and receive medical advice. Uses the responses to recommend one of several possible next steps, e.g. continuing to practice social distancing or contacting a healthcare provider.	Apple's "COVID-19 Screening Tool" [242] CDC's "Symptom Self-Checker" [243]	Symptoms associated with COVID-19 (e.g. fever, cough), risk factors (e.g. age, comorbidities), recent travel, possible exposures.	No. Applications are designed for one-off use and do not collect data longitudinally.
Contact tracing apps	Keeps track of people users have been in contact with by collecting location and proximity data, and alerts users of possible exposures. May utilize a range of technologies, e.g. Bluetooth, GPS, or Google and Apple's joint API.	BlueTrace (Singapore) [244] Corona-Warn-App (Germany) [245]	Collects records of which other devices a device was near over the past days or weeks, enabling alerting of possibly exposed individuals in the case of a positive test.	Yes. Applications track location or proximity information over time.
Longitudinal symptom tracker / contact diary	Helps users track symptoms, possible exposures, and other data over time, potentially enabling discovery of longitudinal trends.	COVID Symptom Study App (UK) [239] StayHome	Symptoms associated with COVID-19 (e.g. fever, cough), risk factors (e.g. age, comorbidities), recent travel, possible exposures. Data is collected repeatedly (e.g. once daily)	Yes. Applications track symptoms and other data over time.

The FHIR specification is flexible in how it can model data and workflows. Supplementary standard operating procedures and constraints are published in Implementation Guides (IGs), supplying additional details and restrictions required for true interoperability, and offering guidance on how to use FHIR to solve particular problems. For example, the US Core IG places restrictions on entity attributes and terminologies (e.g. ICD-10-CM). Additionally, the Structured Data Capture (SDC) IG guides the interoperable and FHIR-compliant implementation of data entry forms. Given the broad nature of the FHIR specification and the existence of mature IGs that address the requirements of the StayHome application, there is an opportunity to use FHIR as the underlying model for both application data and business logic.

7.1.3 Significance

To address the need for community-based self-monitoring and exposure tracking, to support individuals' decision making and contact tracing efforts in case of infection, we developed StayHome, a reusable, mobile-friendly, longitudinal symptom tracker designed and developed following user-centered design principles. StayHome allows regular logging of symptoms and other information, and review of data over time.

StayHome is unique in that it is implemented not just to exchange data using FHIR, but to adopt FHIR resources for internal representation of data and business logic, a design approach we call "FHIR-native". Traditionally, FHIR and its predecessors, such as the HL7v2 messaging standard, are used first and foremost to support an interoperable interface strategy. In line with this primary goal of FHIR, StayHome's use of the standard enables interoperability with other health informatics systems, such as EHR systems and public health informatics systems. However, StayHome also leverages FHIR as internal information architecture, using the standard's domain models to represent both data and business logic. This holistic use of FHIR makes the application a generic PRO tool independent from any specific health problem, PRO use case, or host system.

StayHome is open source and freely available to anyone to use, modify, and implement for clinical and consumer health informatics applications (under the BSD 3-Clause license) at https://github.com/uwcirg/stayhome-client and associated repositories.

With FHIR adoption increasing, informaticists are exploring how best to use it to address important issues in health informatics. We share lessons learned from developing and deploying this application, which we hope will benefit others in designing and developing applications using this new standard.

7.2 Methods

7.2.1 Needs assessment & design

In February 2020, before Washington State's stay-home-orders were put in place and before the outbreak was officially characterized as a pandemic, we recognized the potential benefits of a symptom tracker application. To inform the requirements for such an application, we solicited input from students in an undergraduate class about science, evidence, and health, conducted by one of the authors (WL), where COVID-19 had become a frequent topic of discussion. This process included informal conversations with the students as well as a structured survey asking participants about their COVID-19 concerns, and how a smartphone application might help them with those concerns (HB, WL, SK). Project team members (SK, HB, WL) then tabulated survey results and used content analysis methodology to identify and prioritize user-centered design features. The University of Washington IRB determined on July 6, 2020 that Ethics approval was not required for this project.

With the pandemic actively unfolding, there was an immediate need for the application. We therefore aimed to keep the user experience simple, using standard mobile UI components and interaction patterns where possible, while maintaining a user-centric approach to application design by conducting user tests and revising based on feedback in an iterative fashion.

7.2.2 FHIR

In addition to using FHIR resources to represent our data model, we also made use of the FHIR RESTful API [246] to create and update both metadata (e.g. `Questionnaire` resources) and data (e.g. `Patient` resources) and the FHIR Search API for data retrieval. Concepts from the Workflow Module (e.g. definitional resources) were used for PRO workflow execution. We dynamically encoded logic via FHIRPath expressions as part of questionnaire display items, per the Clinical Reasoning Module and SDC guidance. Internationalization was implemented using guidance from the Implementation Support Module. The `Consent` resource was employed in conjunction with guidance from the Security and Privacy Module to record data sharing preferences. We referenced the Terminology Module to implement a custom code system for app-internal messages/notifications. Extended operations, as described in the Exchange Module, were used to expand answer option value sets.

7.2.3 Development and deployment

Mobile applications exist within several disjoint consumer app ecosystems characterized by different hardware, software, and process constraints, which is a barrier to mobile app development and maintenance. Cross-platform mobile application development frameworks are available to address this challenge, allowing the use of a single codebase for developing Android, iOS, and web applications. This work utilized Google's Flutter [247] cross-platform mobile application development framework.

Figure 7.1 shows an overview of the overall system architecture. In addition to the client applications (community facing and administrative), the system includes Keycloak [248] as an Open ID Connect identity provider, a reverse proxy for role-based access control ("map-api"), and HAPI FHIR version 4.2.0 as a FHIR API server and persistent data store.

Amazon Web Services (AWS) cloud infrastructure provides high availability and scalability for each server component, and Docker supports continuous integration (CI) and

deployment. StayHome is internationalized to English, Spanish, Haitian Creole, and German, with automated string export and import processes.

GitLab and GitHub were used for version control and to make our work publicly accessible. The full code for the StayHome app, the authorization reverse proxy server, and Docker files needed for deployment are open-source and freely available.

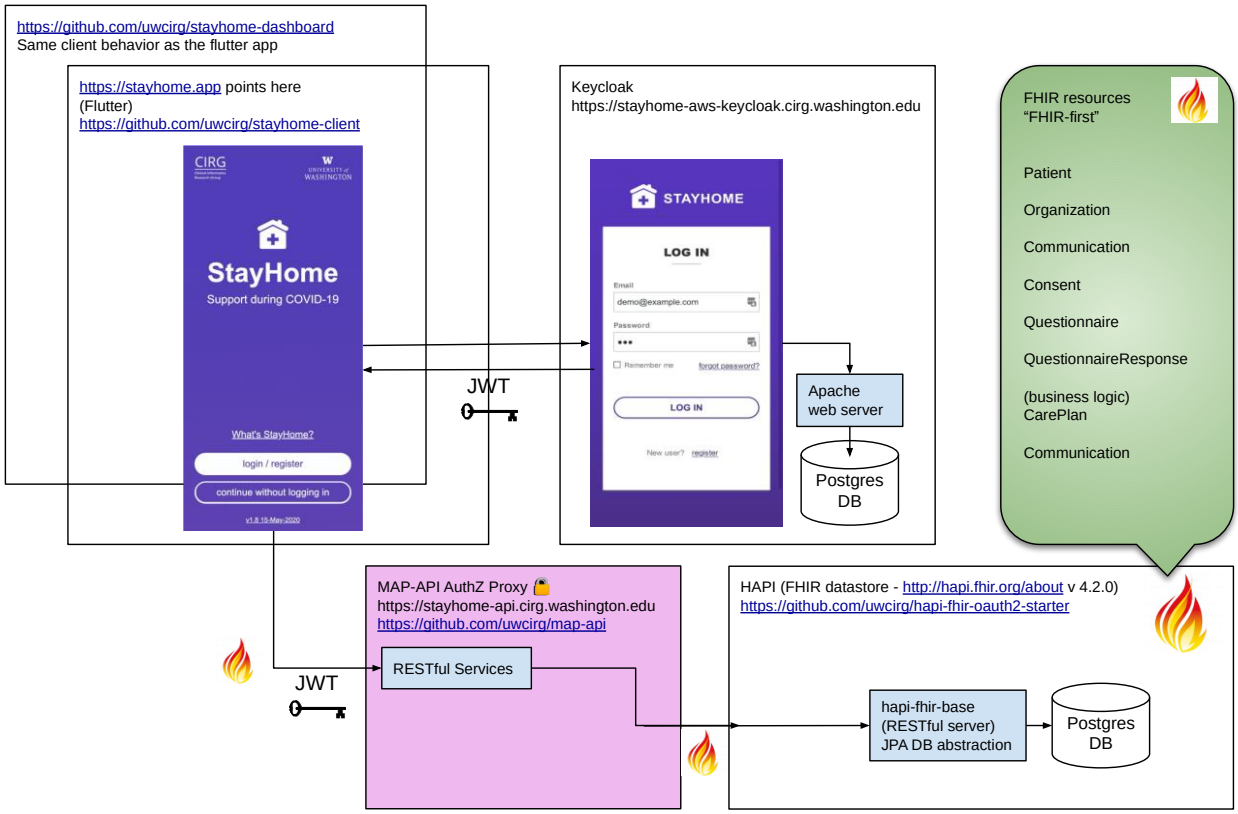

Figure 7.1 StayHome system architecture.

In addition to the client applications (community facing and administrative), the system includes Keycloak as an identity provider, a reverse proxy for role-based access control ("map-api"), and HAPI FHIR version 4.2.0 as a FHIR API server and persistent data store.

We began software development in February 2020 and published version 1.0 on March 27, 2020. StayHome is live and accessible to the public at https://stayhome.app.

A companion dashboard application ("stayhome-dashboard") was developed to allow administrators to view and manage data and users. This application supports public health surveillance and reporting workflows by providing search and filter functionality and allowing review of usage statistics and user level-information such as location (for users who choose to

share that information). The appearance of user data from the client-app is based on “opt-in” permissions given by the user of the client app. Further, the dashboard provides a venue to host future functionality, such as advanced reporting and data visualization.

7.3 Results

7.3.1 Needs assessment & design

There were 63 undergraduate students enrolled in the class; about 40 students were present in class when we conducted the assessment in February 2020. Participants indicated that they were concerned about infection and developing the disease and that they were unsure of how to protect themselves and others. Many were wondering how they would be able to tell if they had contracted the virus. Some participants indicated that they measured their body temperature daily. With word of mouth being a major information source and official sources recommending caution or even de-emphasizing the seriousness of the outbreak, participants expressed uncertainty that fueled a need for reliable information. Based on these conversations with potential users, it became apparent that we had the opportunity to develop an application to address the need for daily self-monitoring of symptoms and clinical signs.

We aimed for a simple, functional design. However, navigation design was challenging and the subject of several design iterations. Our goal was to streamline the user experience so users could quickly find frequent actions, while still allowing easy access to less-frequently used functions. For example, the symptom assessment questionnaire might be completed daily. In contrast, the risk factors and exposures questionnaires would only be filled out once (or in the case of a relevant event). Additionally, we wanted to avoid confusion and provide reassurance in how data would be collected and used, so we opted to display lay-person-friendly terms and conditions and detailed explanations in-line with the controls that asked for such information. Figure 7.2 shows screenshots of the current user interface.

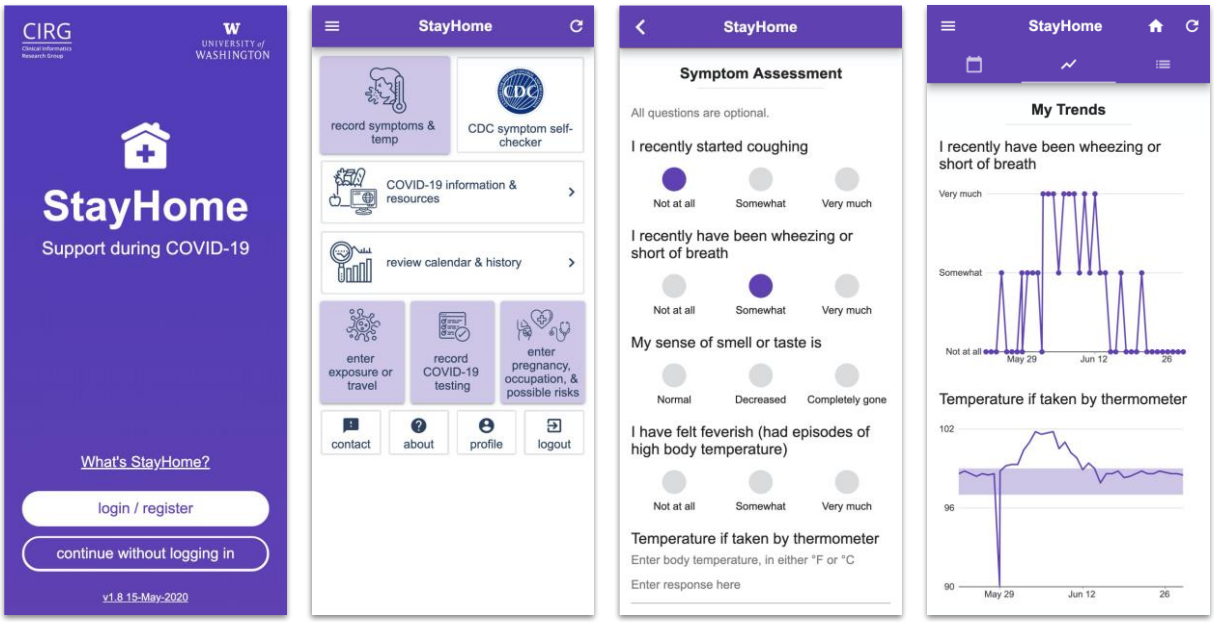

Figure 7.2 StayHome application UI.
From left to right: Login/start screen, home screen, questionnaire screen, trends review screen.

7.3.2 FHIR

The use of FHIR supported the rapid development of the StayHome application by providing a domain-appropriate data model and query framework. FHIR specifies the general structure (i.e. applicable data elements) of resources describing various concepts in health and healthcare applications; there is, however, significant flexibility in how these resources can support individual use cases. This section gives an overview of some of the parts of the FHIR specification used, and how they support the StayHome use case. See Figure 7.3 for an overview of how resources work together.

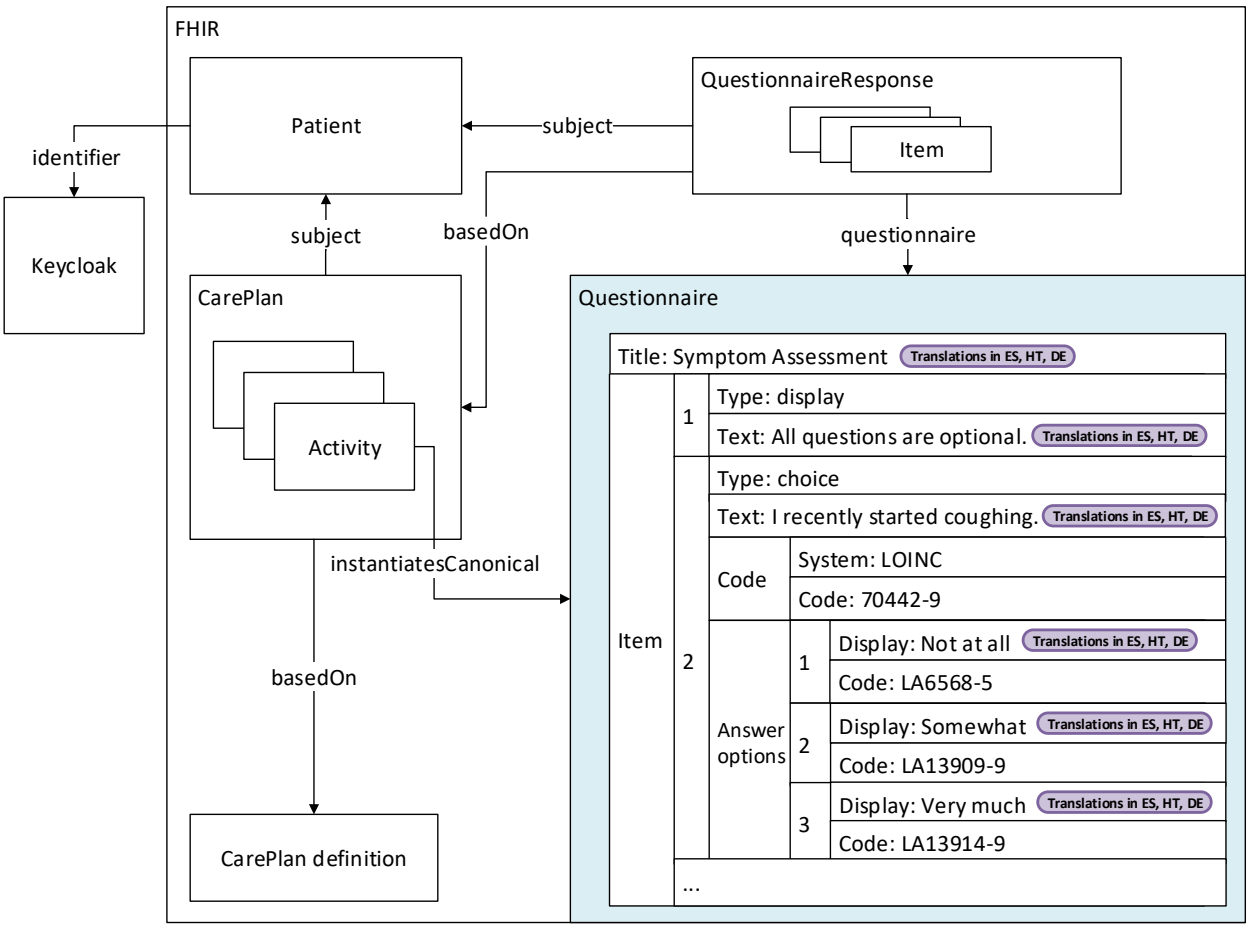

Figure 7.3 FHIR resources used by StayHome.

7.3.2.1 Resources

`Patient`. The patient resource links individual users' personal information and settings with their account login managed by the external identity provider service. StayHome uses `Patient` resources to manage demographics, contact information, and language preference (the application does not require users to enter this information).

`CarePlan`**.** This resource specifies a set of questionnaires. A definitional resource is defined on the application level, which is then instantiated for each new user. Instantiation allows subsequent, person-specific adjustment of the set of questionnaires to be administered and provides the opportunity to personalize the frequency of individual instruments. `CarePlan`

resources reference the patient they belong to via the subject element and list the questionnaires in the activity element.

`Questionnaire`. This resource defines the content of each PRO instrument and is thus central to StayHome. It consists of a list of display items, which can be either an answerable question or help text. Each item includes answer options coded for interoperability (e.g. LOINC codes), selection behavior (e.g. single answer vs. multiple answer), and formatting information (e.g. section headers vs. help text), as appropriate. Items can also represent calculated quantities (e.g. score totals), conditionally displayed items to support branching logic, and conditional user messages (e.g. a "high risk" and a "low risk" message, presented based on the calculated score total). For an excerpt from the symptom assessment questionnaire, see Figure 7.3. `QuestionnaireResponse` resources record responses to questionnaires.

`Consent`. StayHome enables users to indicate their data sharing preferences using `Consent` resources. Each of these resources indicates whether an individual user consented to share their data with the stated organization.

`Organization`. This resource type is used to identify entities to whom sharing consent can be given, e.g. public health agencies, researchers, and potential community partners, and is referenced by `Consent` resources.

7.3.2.2 Query API

StayHome utilizes FHIR's built-in search APIs in several ways. First, resources are queried by ID in cases where that ID is known, e.g. in `CarePlans`, which specify their component `Questionnaires` by ID. Second, the application retrieves relevant resources with simple search queries, e.g. to retrieve the `Patient` resource corresponding to the logged-in user. Finally, multiple combined search and filter criteria are used to retrieve bundles of resources for specific use cases; for example, to plot response history for individual questions. The administrative dashboard uses such a query to retrieve records. The "map-api" reverse proxy ensures that only

resources with proper authorization access are returned, i.e. those belonging to the logged-in user, or on the dashboard, those referring to patients who consented to share data.

7.3.2.3 Internationalization

```
  "text": "Start date",
  "_text": {
    "extension": [
      {
        "url": "http://hl7.org/fhir/StructureDefinition/elementdefinition-
translatable",
        "valueBoolean": true
      },
      {
        "url": "http://hl7.org/fhir/StructureDefinition/translation",
        "extension": [
          {
            "url": "lang",
            "valueCode": "de"
          },
          {
            "url": "content",
            "valueString": "Anfangsdatum"
          }
        ]
      },
      {
        "url": "http://hl7.org/fhir/StructureDefinition/translation",
        "extension": [
          {
            "url": "lang",
            "valueCode": "es"
          },
          {
            "url": "content",
            "valueString": "Fecha de inicio"
          }
        ]
      }
    ]
  },
```

Code snippet 7.1 JSON example of a string element with two translations. The string is tagged as user-facing with the Translatable extension.

To display questionnaires in different languages according to user choice, we utilized both the Translatable and Translation extensions following guidance from the Implementation

Support Module (Internationalization section). The Translatable extension tags strings as user-facing. We added an automatic extraction step to push untranslated English strings to the CI pipeline. After receiving translations from translators working in translation software such as Transifex [249], translations were inserted back into the FHIR resources using the Translation extension (Code snippet 7.1).

7.3.2.4 Calculated Questionnaire Items

The FHIR Structured Data Capture (SDC) IG describes how to use expressions and extensions to support calculated display items, such as score totals. StayHome employs FHIRPath as an expression language, and the ordinalValue extension attaches numeric values to categorical response options. The freely available fhirpath.js [250] library is used to calculate the value of the expressions in real-time as the user enters questionnaire responses. The `Questionnaire` resource includes the expression shown in Code snippet 7.2 for this purpose.

```
QuestionnaireResponse.item.answer.extension.where(url.contains('valueOrdina
l')).valueDecimal.aggregate($this + $total, 0)
```

Code snippet 7.2 FHIRPath expression used by StayHome to calculate the sum of all selected answers' ordinal values (score total).

7.4 Discussion

In this work, we designed, developed, and deployed a mobile-friendly, FHIR-native, PRO tool with application to COVID-19 symptom assessment and exposure diary functions. This information system administers questionnaires and organizes information using FHIR for the representation of both application content and behavior.

7.4.1 FHIR-native approach

7.4.1.1 Advantages

Using FHIR in general and in a FHIR-native application context has significant advantages (Table 7.2). The FHIR standard emerged from the modeling work of many experts

and thus provides an excellent starting point for an application-specific data model in the health domain. While our application's needs are comparatively simple, starting with data structures capable of accommodating advanced data and process models obviated the need to restructure the data model in the face of new business requirements. Existing FHIR server implementations such as HAPI provide routine functionality, such as create, read, update, and delete (CRUD) APIs, as well as advanced domain-specific tooling, such as CQL support, allowing developers to focus on their application's unique requirements. Because FHIR specifies the interface between clients and servers, implementations are interchangeable as long as they comply with this interface. This allows the operation of apps directly against either standalone FHIR servers or other FHIR-compliant servers, e.g. those belonging to EHRs or public health information systems. FHIR also facilitates the exchange of information between systems, known as interoperability, though carefully following implementation guides is required to achieve semantic interoperability.

Beyond data exchange, we leveraged FHIR as the internal information architecture. This approach facilitated rapid application development by reducing the need to design and refine a data model, as well as the tooling to support it. Because FHIR resources (e.g. definitional resources and questionnaires) drive application content and behavior, PRO functionality is reusable for any FHIR-compliant questionnaire. The application requires no custom server code and operates entirely within the FHIR specification, facilitating the reuse of StayHome as a generic FHIR-enabled questionnaire administration tool. FHIR can accommodate complex business logic in an elegant, standardized way, making possible a wide range of applications that require no server code.

7.4.1.2 Disadvantages

There are disadvantages to using FHIR, which may intensify for FHIR-native applications. Though the ability to use existing server implementations and tools offsets the

time investment required to use FHIR, the complexity and size of the FHIR specification and its extended documents may be a barrier to adoption. Additionally, in some domains, e.g. in the consumer health application realm, FHIR may not be widely adopted, reducing the immediate benefit of interoperability-readiness. Further, many use cases, clinical and non-clinical, are not modeled by FHIR. While the standard is flexible and extensible to new scenarios, there is a risk of tipping the balance from using it appropriately to counterproductively shoehorning application data and business logic into FHIR. For example, if resources or resource elements cannot be understood outside of the StayHome context or are not compatible with recommendations laid out in commonly used IGs such as US Core, the use of FHIR may turn out to be a burden.

Finally, despite the guidance from FHIR modules and IGs, many implementation and representation choices remain with the developer, and these design choices may be imperfect. For example, we considered using a single, unmodifiable `CarePlan` resource to define the set of questionnaires at the application level. We instead chose to instantiate a user-level `CarePlan` resource with reference to an application-level definitional resource (template) to allow personalization of app content. This instantiation introduced additional barriers for content maintenance: adding new questionnaires to the definitional resource required updating every user's `CarePlan` instance as well (previous versions are still accessible via the history API). This experience underscored the need to make representational choices carefully and with specific use cases in mind; to ground decisions in implementation guides, which address many common difficulties, as much as possible; and to accept the fact that updating large numbers of resources as part of application updates may ultimately be unavoidable as business requirements evolve.

Table 7.2 FHIR/FHIR-native advantages and disadvantages.

Pros	Cons
– Represents modeling work of many experts – Specifies common requirements and corresponding tools/APIs (e.g. storage functionality, queries, versioning, CRUD operations, data formats, localization) – Specifies advanced functionality; specific implementations (e.g. HAPI) might provide further functions that become automatically available to FHIR-native apps (e.g. Async to export bulk data, terminology services to translate codes, advanced searches, subscriptions, CQL) – In theory, EHR workflow integration out of the box with SMART [251] – In principle, standard data format enables immediate interoperability with other systems: Resources can be consumed by other apps and other apps' resources could be consumed by StayHome – IGs help with solving common challenges and further semantic interoperability – FHIR implementations are interchangeable (e.g. HAPI, Azure and Google Cloud FHIR servers, EHR with FHIR APIs, public health system with FHIR APIs) – Clients are independent of the server / can operate out of any FHIR-compliant server (e.g. a standalone system or an EHR). – PRO functionality is reusable to any questionnaire that can be specified as FHIR – Built-in backward compatibility (between the data model and different versions of client apps) – Standard terminologies enable translating the application to work in a new context (e.g. to use LOINC instead of SNOMED)	– Complex: it can be challenging to determine the intent of different pieces, and what applies to a particular use case; takes time to work through; developing compliant applications requires consideration of constraints from different places – Specific application constraints and requirements may not have been considered when developing the spec – Does not specify all functionality required for a complete app, e.g. access control, workflows for non-clinical use cases – FHIR does not model some specialized/complex clinical use cases – Risk of forcing data and business logic into FHIR in a way it wasn't intended for, compromising interoperability and application reusability in practice – May currently be limited in some application domains, e.g. wearables and other consumer health applications

7.4.2 Development and deployment

For our user-facing application, we chose Flutter as a solution to cross-platform mobile application development, but quickly felt the impact of using emerging technologies. For

example, there was no FHIR resource library for the Dart programming language, requiring us to generate and test the data model and serialization code. On the other hand, using an actively evolving framework enabled us to take advantage of cutting-edge technology. For example, Flutter Web was first released in the beta channel in December 2019. Not long before our initial release, we recognized that our group did not have the necessary resources and expertise to overcome the challenges of mobile app publishing (e.g. registering as a non-profit entity, integrating iOS application builds in our CI pipeline), within an acceptable time frame. Flutter Web, while suffering idiosyncrasies typical of beta-stage software, was an alternative that became available just in time for us to publish StayHome as a web application instead.

Web applications have many benefits over native applications. They provide the same experience on different platforms, including desktop, maximizing reach while minimizing the additional development effort required to support many different platforms. Deployments (such as when a new version is published) affect end-users minimally, as client browser apps will automatically retrieve the latest version of the web application when reloading, eliminating the need to update local application installs. Additionally, some development teams may find deploying a web application to be straightforward compared to completing the steps required to get a native mobile application approved and published on each of the major platform app stores. While web applications may provide suboptimal UX compared to native apps, there is an opportunity to narrow that gap via Progressive Web Application (PWA) or "installable web application" setups. Other drawbacks of non-native applications include limitations for integration with local hardware (e.g. Bluetooth, camera) and barriers in using notifications.

Authentication and authorization for FHIR were challenges, as fine-grained data access control is an infrastructure requirement not adequately addressed by FHIR. While solutions such as SMART-on-FHIR and Interceptors for HAPI FHIR were a possibility for authentication and authorization, we instead implemented a lightweight reverse proxy server as a single solution for both authentication and authorization. This approach removed the need to manage

and secure credentials within the StayHome app and allowed us to keep the server implementation (e.g. HAPI vs. EHR) interchangeable and independent of the client app. While comparatively simple conceptually and in practice, this “non-standard SMART-on-FHIR” setup adds a custom layer in our system architecture. We are actively investigating the use of a so-called standalone SMART-on-FHIR application setup to simplify our authorization system.

7.4.3 Future work

Future work includes implementing a fully SMART compliant launch flow. Additionally, there is an opportunity to further leverage FHIR to support more advanced functionality, for example, by using extraction as described in the SDC module to create `Observation` resources from `QuestionnaireResponse` resources.

While we provided the application for use to members of our immediate community to support community members as soon as possible and to collect further feedback, accomplishing broad adoption of the app requires resources and expertise, including marketing and consumer support, that lie outside of our capabilities and resources. However, the application may be useful for specific programs, such as research projects and community health programs, with well-defined objectives and user groups. Such small- to medium-scale implementations thus represent an important future direction for this work.

7.5 Conclusion

We presented StayHome, a FHIR-native PRO tool applied as a COVID-19 symptom tracking application and public health reporting tool. We deployed this tool in a software system that includes a community member-facing web application, authentication/authorization layer, FHIR server, and administrative dashboard application. Code and resources are open source under the terms of the BSD 3-Clause license. As a mobile-friendly web application, StayHome maximizes reach. StayHome addresses core community needs by providing users with a tool to

self-track possible symptoms and exposures and supports them in following public health recommendations for their personal health. The use of FHIR enables user-controlled interoperability with, for example, public health contact tracing and case investigation systems, an area the authors are actively working on with the Washington State Department of Health through the CommonCircle initiative. We innovatively operationalized FHIR beyond its use as a messaging standard, leveraging it as an internal information architecture to represent application content and behavior. Thus, we created a questionnaire administration tool that is easily reusable for other use cases by simply updating FHIR resources. The tool can operate out of any host system, e.g. a standalone FHIR server, an EHR, or another FHIR-compliant server. During this process, we found that there are advantages and disadvantages in the use of FHIR and the FHIR-native approach. We applied FHIR in a novel way by using it as internal information architecture, but also found that the right level of FHIR use can be a balancing act that depends on individual application requirements and constraints. We used emerging technologies, tried-and-true technologies, and custom system components, finding benefits and tradeoffs for each. We also recognized that pre-existing expertise is a factor in deployment considerations. We provide our code in several open source repositories, freely available under the BSD 3-Clause license at https://github.com/uwcirg/stayhome-client and in associated repositories. We hope that the StayHome application benefits the community in working to stay safe and flatten the curve. Additionally, we hope that the code and resources found in our Github repositories and the considerations shared in this manuscript benefit others in their efforts to develop and operationalize FHIR applications in general and PRO applications in particular.

7.6 Acknowledgements

The authors conducted this work through the UW Clinical Informatics Research Group (CIRG), and gratefully acknowledge the contributions of CIRG staff, especially Justin McReynolds.

This work was partially supported by the National Library of Medicine Training Grant (T15LM007442).

Chapter 8. A FHIR-based approach to text message-based suicide prevention: Informatics-Supported Administration of Caring Contacts (ISACC)

In addition to opportunities for automation and cognitive support, aim 1 (Chapter 3) revealed that data and workflow being spread across multiple different systems poses a host of challenges. Consequently, a core design focus for informatics tools for Caring Contacts should be to enable data and workflow integration. Further, in aim 2, I developed a prioritization model of suicide risk, which has the potential to facilitate scalable deployment of Caring Contacts when leveraged for message triage. In the work described in this chapter, I therefore investigated how automation and cognitive support can be provided in an interoperable, workflow-integrated way. Specifically, considering that health data standards FHIR and SMART-on-FHIR provide mechanisms for interoperability and portability, I developed a FHIR-native data representation model for text message-based suicide prevention and a SMART-on-FHIR application architecture capable of utilizing my prioritization model for real-time triage. Finally, I implemented the Informatics-Supported Administration of Caring Contacts (ISACC) tool, a standards-based, open-source, AI-supported information system for Caring Contacts, demonstrating the feasibility of using FHIR and SMART-on-FHIR for this purpose.

Abstract. Background: Suicide is a leading cause of death in the United States. Caring Contacts, a text message-based intervention, is effective in reducing suicide, attempts, and ideation, but adoption is hampered by a lack of appropriate technological support. Achieving interoperability for patient-generated data (PGD), such as text messages, to enable their use within clinical workflows is difficult because of diverging practices between consumer and clinical technologies. Similarly, ensuring portability and reusability of clinical decision support

(CDS) across systems is challenging due in part to variations in data formats and communication protocols across health information systems. Having seen broad acceptance and adoption, current data and messaging standards such as Fast Healthcare Interoperability Resources (FHIR) enable seamless data exchange and software re-use. However, there are few reports on how to use FHIR to represent PGD that is continuously generated and available only from sources external to the EHR, or how to architect FHIR-based systems to ingest this data and process it for CDS in real time.

Objectives: To design, develop, and share a FHIR-based, interoperable, portable, open-source software system for suicide prevention that uses patient text messages to provide CDS within clinical workflows.

Methods: Based on a needs assessment conducted with suicide preventionists representing a range of healthcare organizations and clinical settings, we developed a list of possible solutions to meet these needs. We then determined application requirements for an initial version of the tool (minimum viable product, MVP), assessed the corresponding data model requirements, and designed a FHIR representation model that meets these requirements. We developed a system architecture capable of supporting the required data artifacts and use cases, including real-time processing of PGD with machine learning, and developed a freely available software artifact embodying these principles.

Results: The Informatics-Supported Administration of Caring Contacts (ISACC) system consists of a web client app and a server application continuously monitoring for and processing SMS communications via machine learning, both of which use SMART-on-FHIR to authenticate against the FHIR API host system (e.g. Electronic Health Records (EHR) system). FHIR is used to represent data and business logic as well as for communication; externally generated data is transformed into FHIR representations for this purpose. All transactions between client, server, and host are conducted via FHIR APIs, so the system can be supported by any host system implementing the appropriate FHIR APIs. The system has the capability to make use of two

different natural language processing (NLP)-based machine learning models, one for assigning suicide risk scores to patient-generated messages, and one for extracting clinically actionable concepts from these messages. The system invokes the models in real-time and presents inferences as part of the workflow. The system leverages FHIR data structures, APIs, and the SMART-on-FHIR integration protocol in a way that both supports the intended use cases and corresponds to best. The software developed as part of this work is freely available from https://github.com/uwcirg/isacc-environments and linked repositories.

Conclusion: We demonstrated the feasibility of using FHIR and SMART-on-FHIR as the basis for a suicide prevention CDS system that leverages text PGD and NLP to support clinical workflows, enabling out-of-the-box interoperability and portability. The system serves as a blueprint for a range of systems seeking to leverage PGD in clinical workflows. Code is freely available to use, adapt, and implement under the BSD 3-Clause license. Next steps include a pilot implementation of the tool at several clinics.

8.1 Introduction

Caring Contacts is an effective suicide prevention intervention, but avenues for more efficient delivery must be explored before it can be widely adopted. Across three large randomized trials, Comtois et al. [40,49,50] refined Caring Contacts via text message and, based on expert consensus, collected an initial list of challenges and bottlenecks that a digital solution may help address. In Chapter 3, I expanded and refined this list and developed design considerations for a digital tool supporting the Caring Contacts intervention [221]. Key requirements included the automation of labor-intensive workflow components, providing cognitive support at appropriate points in the workflow, and enabling seamless data and workflow integration. Specifically, tools that help providers manage text message streams as part of the electronic delivery of Caring Contacts, including artificial intelligence (AI)-based clinical decision support (CDS) to prioritize messages based on content and support staff in

composing follow-up messages, are needed. In Chapter 4, Chapter 5, and Chapter 6, I explored avenues for such cognitive support.

Our finding of the importance of data and workflow integration confirms long-standing guidance for effective CDS, which emphasizes the need to avoid adding additional workflow steps and to provide information at the right time, to the right person; and to provide the right information, i.e. information that is actionable [162,252]. Fortunately, we now have technologies that can be harnessed to meet these needs. Data and workflow integration standards and protocols are available and have seen broad adoption in recent years. An internationally recognized standards specification for health-related data structures and exchange protocols, Fast Healthcare Interoperability Resources (FHIR) [26] is now mandated as part of the 21st Century Cures act [145,146] requiring U.S. health information technology vendors to enable interoperability. Technologies such as SMART-on-FHIR [25], an authentication and integration framework for FHIR-based health apps, are available to develop such applications and integrate them seamlessly into clinical workflows. However, it is not clear how these technologies can be harnessed in conjunction with AI to provide automation and cognitive support in an interoperable, workflow-integrated way in the context of the Caring Contacts suicide prevention intervention.

8.1.1 Contributions of this work

In this work, we designed and developed a FHIR-based information system for Caring Contacts. The system incorporates patient-generated data (PGD) and leverages machine learning for CDS. We outline the data model requirements for this application, and present how FHIR can be used to meet these requirements. We present detailed documentation of how to use FHIR for this application area. Further, we make the code freely available under the BSD 3-Clause license. Using Shaw's categories of software engineering research contributions [253], our contributions are as follows:

1. A novel, FHIR-native data representation model for AI-supported, text message-based suicide prevention interventions
2. An architectural style/design pattern for applications reading data from sources external to electronic health record (EHR) systems, processing them in real-time with AI to extract clinically actionable signal, and integrating these insights into clinical workflows
3. An implemented tool (open-source software artifact) that embodies 1 and 2

8.2 Related work

8.2.1 FHIR & CDS with SMART-on-FHIR

The landscape of healthcare information technology is fragmented for a number of reasons. Health data are created in many different settings, and medical sub-domains have developed their own approaches to storing, processing, and reusing these data. Health information technology (HIT) vendors have independently developed purpose-built solutions, resulting in siloed systems with proprietary data structures and data exchange strategies; consequently, data exchange between systems is difficult, requiring case-by-case interface solutions, and driving up integration costs. This also results in systems not being easily portable between settings, i.e. informaticians may need to undertake substantial customizations in order to deploy their apps at healthcare organizations with different technology ecosystems. If systems are not interoperable and workflows are not seamlessly integrated, the user experience suffers, as clinical users must navigate many different systems and duplicate data entry efforts. These inefficiencies severely affect the usefulness of HIT. Additionally, the vision of the Learning Healthcare System [8,10] can only be achieved if data from different sources can be collected and harmonized for secondary use, i.e. analytics. Cost-effectiveness, usability, and sustainability considerations are therefore key components of developing health information systems.

This has led to the increased recognition of the importance of standards-based HIT in recent years. Health information technology data standards have a rich history, most recently

culminating in international standards organization Health Level Seven International (HL7) releasing FHIR [26]. FHIR was designed to address many of the shortcomings of its predecessors that caused low adoption, such as a failure to make use of contemporary technologies, e.g. XML or JSON, in favor of the less pervasive RDF format [25]. FHIR is a specification of the structure of resources, such as Patient and Observation resources, which represent clinical, administrative, and infrastructure-related concepts. However, FHIR also specifies application programmer interface (API) endpoints, which constitute ways to interact with a database server that deals in these resources, using HTTP. For example, it specifies how patient records can be queried by RESTful web services, by fields such as phone number. FHIR is now broadly adopted, with the major EHR vendors implementing FHIR APIs [159]; however, it is important to note that vendors are at different stages in the implementation process. Additionally, the implementation of write-APIs is limited by important questions of data governance and quality, which must be carefully considered before APIs can be made available to third parties.

Substitutable Medical Applications and Reusable Technologies (SMART) on FHIR [25,159] is a framework for authentication and authorization of apps seeking to consume FHIR resources from a FHIR-enabled host system, such as an EHR implementing FHIR. It provides a way to launch apps within this host system, transferring the launch context to the launched app, such that end users need not re-authenticate, re-select the appropriate patient chart, or repeatedly specify which functionality they want to access. SMART-on-FHIR apps can be written by independent app developers who have complete control over business logic and user interface, while allowing the appropriate use of FHIR resources internal to the host system. In this way, it enables completely customizable CDS functions without needing the EHR vendor to specifically support each app, and without disrupting the user workflow.

8.2.2 Use of FHIR for PGD

PGD provide a unique opportunity to enhance clinical care. Rather than being created by doctors and other healthcare workers as part of clinical care, PGD originate with the patient. They span many types of data including patient-reported outcomes data (PROs), such as experience data collected via questionnaires; consumer health device data, such as from wearables or Wi-Fi-enabled bathroom scales; and many other types of data created by and/or belonging with the patient, such as smartphone location logs and social media posts [82].

Leveraging PGD for health-related applications is a promising frontier. PGD contains information that cannot be found elsewhere in the EHR, and may have unprecedented completeness and quality, particularly if automatically captured [87]. As a result, this has been a fruitful area of research. Numerous examples exist in the machine learning literature of using PGD for making health-related inferences, e.g. for predicting depression and anxiety from location and physical activity data [254] or from social media data [123,205].

However, integrating PGD and PGD-derived insights into clinical care raises many unique challenges. Besides questions of patient privacy, PGD is difficult to get into clinicians' hands for reasons of interoperability and other technical challenges [84]. Various approaches have been used to overcome these challenges. Some have used patient portals to circumvent interoperability and security/privacy challenges [255], but using patient portals may not be possible or optimal in some cases for a range of reasons. Since 2011, Sim and colleagues have been developing Open mHealth [256,257], a framework to standardize PGD such as sensor data from wearables. Open mHealth includes schemas for a wide range of patient-generated health data, such as activity, body weight, sleep, and blood pressure data. While the Open mHealth standard is not inherently interoperable with EHR systems, OMH on FHIR [258] provides a pathway to converting this structured patient-generated data into FHIR resources. Torous et al. [259] developed a system for capturing and integrating PGD from patients; however, this system is intended to be a comprehensive PGD system capable of handling high volumes of raw sensor

data, and thus the authors found FHIR too limiting for two reasons: first, the lack of semantic interoperability standards for PGD (i.e. there is no broadly accepted Implementation Guide, such as US Core [260] for other data), and second, the lack of write access to FHIR APIs implemented by prevailing EHR vendors. This led Torous and colleagues to use proprietary APIs and data structures to fulfill their specific needs. Finally, Apple's HealthKit is a powerful framework for consumers to collect their health data, including their clinical health record data, in a centralized location; it uses FHIR to obtain records from healthcare organizations [261]. However, integration of PGD from HealthKit back into the EHR requires vendor-specific solutions; for example, Epic integration is mediated by the patient portal application MyChart [262]. Other reports in the literature describe applications conceptualized and designed as FHIR-based applications, but have not been implemented; others were never intended to automatically and continuously ingest PGD [85,263].

Integrating natural language data has further unique challenges. For example, in case a patient chooses to share natural language data such as social media posts, diary entries, or text message logs, the usefulness of raw data is significantly limited. Even with structured data, the volume of data and low likelihood of clinical relevance limit their clinical usefulness [85]; while numeric data can be plotted over time or checked against defined normal ranges, this is not possible with free text data because it is not inherently computable. Unsurprisingly, a 2020 systematic review on PGD and EHR integration found no studies using patient-generated natural language data [85].

8.2.3 FHIR and text messages

There are few published examples of FHIR being used to model text messages.

Bass et al. [264] developed Configurable Assessment Messaging Platform for Interventions (CAMPI), a platform enabling communication between health intervention teams and patients via SMS. In one implementation, CAMPI was used to deliver text messages

containing intervention content on a pre-programmed schedule with tapering frequency - first daily, then weekly, and finally monthly. In another, the system was used to collect participants' self-reported stress levels, with participants being able to text back their responses. CAMPI uses a message templating approach, where messages can be specified with placeholders, e.g. "Hi [name], Please complete the following survey…". A delivery timeframe is specified for scheduled messages, and the system keeps carefully manages message delivery status to avoid sending duplicated messages, which is required because of the potentially asynchronous nature of text message delivery channels. Though CAMPI does not make use of FHIR, their use case aligns well with ours, and the authors suggest FHIR as an avenue for future work exploring EHR integration.

As part of their extensive library of technical frameworks, Integrating the Healthcare Enterprise (IHE) International published a trial implementation FHIR Profile titled Mobile Alert Communication Management (mACM), intended to guide the implementation of basic alerting services with a low barrier to entry [265]. The profile details the use of FHIR resources, specifically Communication and CommunicationRequest, to model text message alerts for two use cases: alerting health workers to a crisis, and for patient care reminders (e.g. for upcoming appointments or medication reminders) intended for patients and caregivers. The authors envision a transaction pattern involving two actors: the alert aggregator, which holds the details of alerts to be disseminated and manages them, including dispatching them to a communications platform for delivery; and the alert reporter, who originates or relays alerts, translating them in order to make them interoperable with the alert aggregator. Further, the profile specifies additional constraints on CommunicationRequest and Communication resources, such as changing the cardinality of several attributes such that at least one entry is required (e.g. the cardinality of the category attribute in CommunicationRequest resources is changed from [0..*] to [1..*], with the coding marking the resources appropriately for dissemination by the alert aggregator). This FHIR report/profile is intended to cover a broad

range of use cases and is concordantly general; here, we follow these recommendations where appropriate and applicable both for application transaction pattern (and therefore, architecture design) and for the FHIR data representation model design.

8.2.4 FHIR for representing machine learning inferences

FHIR has been used in conjunction with machine learning applications. For example, Hong et al. [266] developed a natural language processing (NLP) pipeline that extracts entities from clinical notes, and generates the relevant FHIR resources, such as `Condition` and `Procedure` resources. For example, the pipeline might recognize mentions of hypertension and hypercholesterolemia (and associated attributes such as onset date and status) in a clinical note, and create a Condition resource for each. The intention is to alleviate the duplicated effort of describing conditions in a note and also adding them as discrete entries on the problem list. FHIR extensions were defined to characterize the NLP origin of the resources; for example, the name and version of the NLP system that extracted the term, and the line and offset in the input note that corresponds to the entity. FHIR extensions have advantages and disadvantages: they allow extending the data model with properties of any name or type, resulting in near unlimited flexibility; at the same time, this exacerbates issues of semantic interoperability, which requires that data structures and the range of possible values are well-defined. Further, the authors conceptualize NLP outputs in terms of standard value sets; for example, an extracted condition's `Condition.clinicalStatus` should be framed in terms of the HL7 value set `ConditionClinicalStatusCodes`. The authors additionally note the necessity of combining existing structured EHR data with the NLP outputs in order to create sufficiently contextualized FHIR resources; for example, inferences may be attached to existing resources, or may reference existing resources such as a `Patient` resource. The authors populate some attributes according to metadata from the source document, e.g. the type of the source document may be used to

populate the `Condition.category` attribute in terms of `ConditionCategoryCodes`, and the `Condition.recordedDate` may match the source document's `Condition.recordedDate`.

8.2.5 FHIR-native

FHIR enables interoperability between health information technology systems with different underlying data representation models by specifying the formats and APIs used for data exchange. However, unlike its predecessors, such as the HL7v2 messaging standard, FHIR provides much more than an interface strategy. The specification consists of a comprehensive representation model for common data types in the clinical domain. A community-powered standard, FHIR is the product of the data modeling efforts of numerous clinical and technical experts. Additionally, it specifies schemes for common transactions, such as creating, reading, updating, and deleting (CRUD) data, as well as for more complex tasks such as querying and transforming resources. It is therefore a suitable starting point for any application working with clinical data.

We previously developed an approach to FHIR application data modeling termed FHIR-native [23]. In the FHIR-native approach, FHIR is the application's primary underlying information model, representing data, state, and business logic. This removes the need for ancillary, non-FHIR database systems and allows us to leverage existing tools, such as open-source reference implementations, to their full extent. In settings where apps operate out of an EHR system, the EHR's database is used as the primary data store.

This approach may not be appropriate for complex applications or use cases not covered by the FHIR standard. FHIR is designed to cover 80% of the use cases in the medical domain; in other words, resources are intended to meet the most common needs [267,268]. As a result, FHIR applications for certain specialized use cases may have data modeling requirements for which FHIR is not appropriate. Applications for these use uses require an additional data representation model beyond FHIR. However, where its application is possible, FHIR-native

can facilitate rapid development and support system scalability and affordability by maximizing the reuse of existing knowledge and tools and minimizing the need for custom solutions and system components.

8.3 Methods

8.3.1 Application requirements

We previously conducted a needs assessment in a purposive sample of 16 interview participants [221]. 11 (69%) participants were invention staff on a Caring Contacts research study, and 5 (31%) worked with programs delivered as part of ongoing care. 11 (69%) participations used text messages to deliver the intervention. Based on our previously reported needs assessment, we developed a model of the current workflow (Figure 8.1). There are three primary actors in the work system: the patient, clinical Caring Contacts team members, and clerical team members. In real-world settings, the same staff, e.g. social workers, often fulfill both roles; however, there have also been implementations where an assistant takes on clerical tasks and a clinician, e.g. behavioral health provider, fulfills the clinical role.

The first actor in the work system is the patient. Patients receive messages on a pre-determined schedule via the designated communication channel, based on patient preference or the healthcare organization's capabilities. For example, messages may be sent via text message, email, or postal mail. Patients may or may not reply to these scheduled messages; if they do, they may then receive follow-up communication from the Caring Contacts team.

Clinical roles complete a range of tasks requiring clinical judgment. For example, they may monitor incoming messages, triage them and decide how to follow up, and also complete the follow-up. In writing follow-up messages, clinical staff must ensure compliance with the "spirit of Caring Contacts", i.e. messages should be caring, undemanding, and appropriate for the patient's situation and needs. A range of information may be referenced to complete this task, including intervention-specific records such as intake notes summarizing patient

information and intervention strategy; patient information from an EHR, if available, including the patient's suicide safety plan, care team, diagnoses, appointments, and history of depression or suicidality score questionnaires; and the message exchange history, which might provide insight into the patient's communication style and baseline risk level. Clinical staff may also create appropriate documentation of intervention midpoints or events, such as clinical notes intended for the patient's primary care or behavioral health provider.

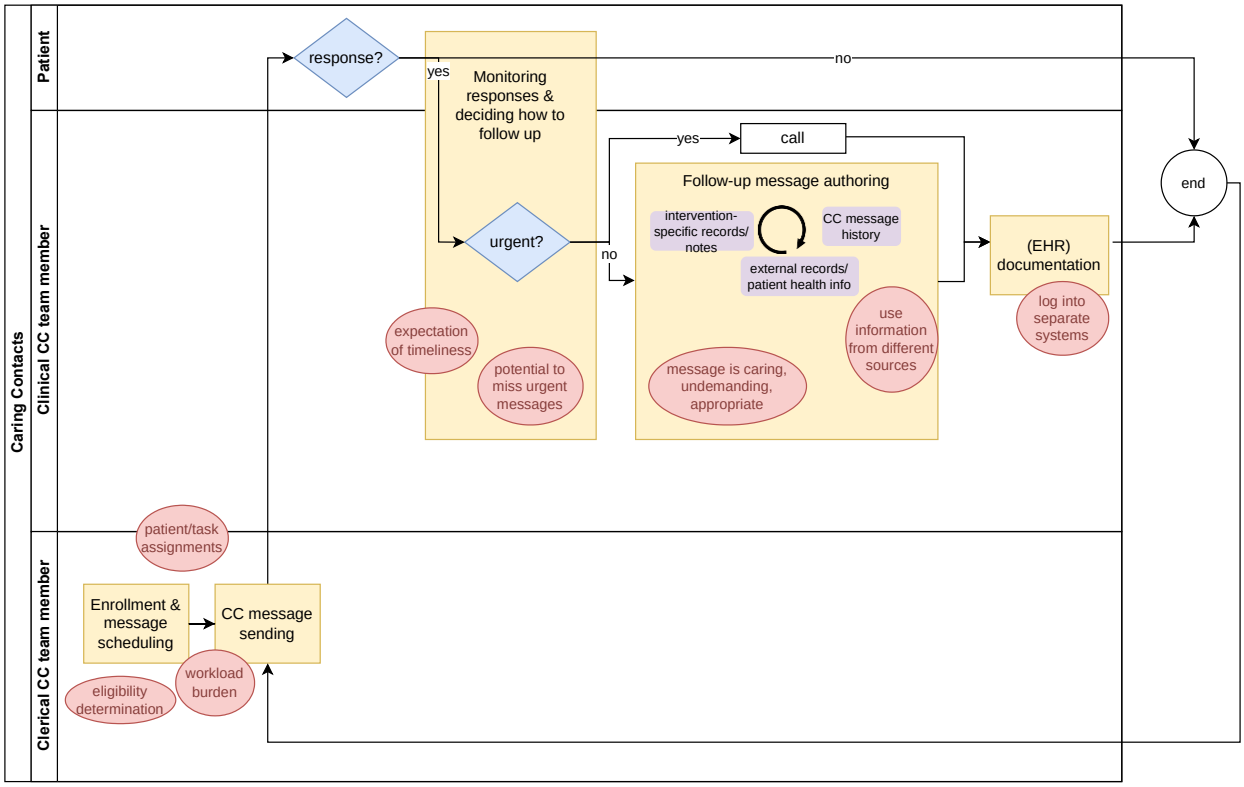

Figure 8.1. Current workflow & actors (without ISACC). Labels A-E correspond to elements of workflow.

Unlike clinical tasks, clerical tasks require little to no clinical expertise. While clinicians may complete these tasks in settings without clerical support, shifting them to assistant staff or information technology can free up clinicians for clinical work. Clerical work includes administrative and logistic tasks, such as determining eligibility based on predefined criteria, enrolling participants and entering their data in the appropriate databases, and scheduling the Caring Contacts messages. They also keep track of patients and their message schedules over

time, e.g. by checking the patient list daily to determine who should receive a message that day, printing cards, preparing postal mailings, and creating emails and text messages as appropriate.

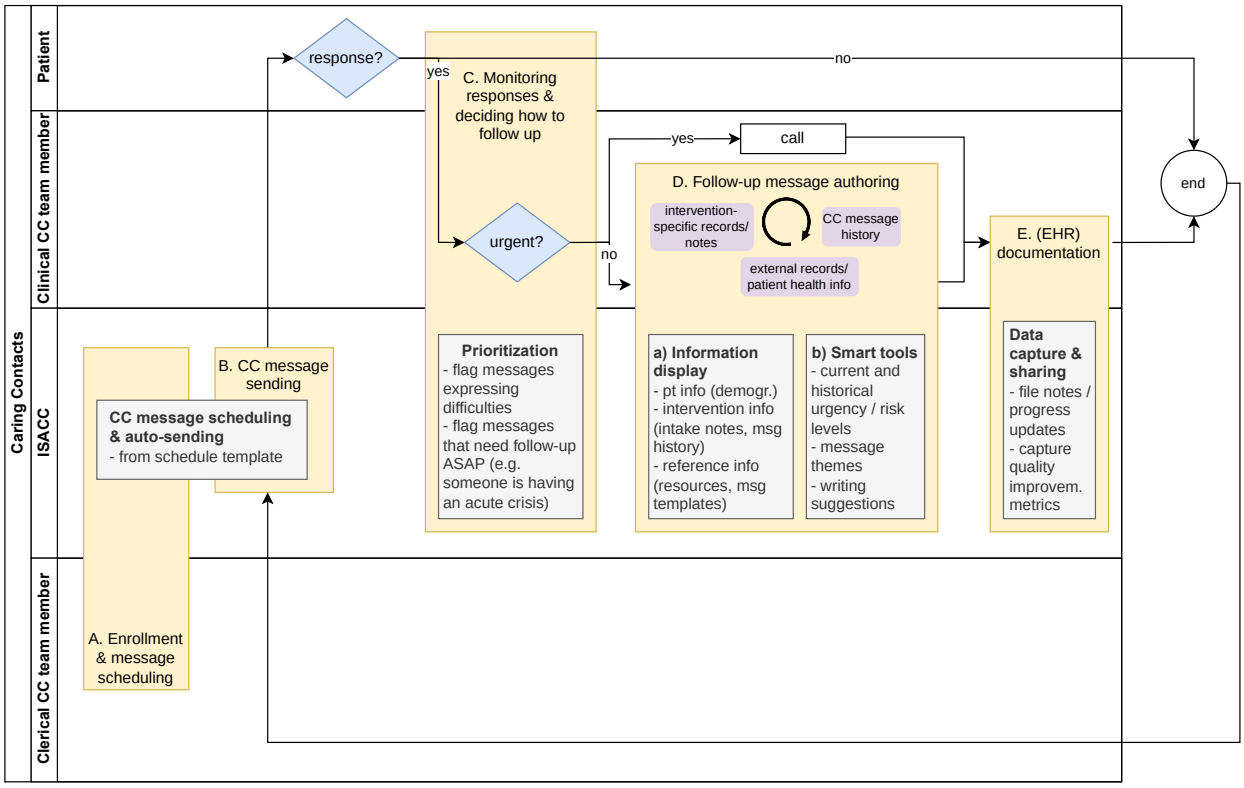

Figure 8.2. Proposed workflow & actors (with ISACC). Labels A-E correspond to elements of workflow.

In our prior work [221], we identified three ways in which technology could support and augment this workflow: automation, cognitive support, and data and workflow integration. The need for automation arises from the workload demands of the intervention combined with resource scarcity. Currently (Figure 8.1), intervention staff spend significant amounts of time on manual, time-consuming tasks, such as keeping track of message schedules, manually creating and sending messages, and monitoring incoming messages to ensure no urgent calls for help are missed; at the same time prevention efforts are often not reimbursed by payors, constraining the amount of time that can be dedicated to such programs. A second area of opportunity for informatics tools is cognitive support for clinical decision making, e.g. with informative dashboards or automatic risk assessment algorithms, which would further free up staff to

support patients rather than wasting time on busywork. Finally, the need for improved data and workflow integration in current workflows is evident from duplicated data entry efforts and inefficient workarounds for communication between stakeholders such as behavioral health providers, social workers, and administrative assistants.

Figure 8.2 shows features informatics tools may implement to realize these design considerations to achieve a re-distribution of tasks and responsibilities. ISACC is depicted as the fourth actor in the work system, which interacts with the patient, clinician, and clerical actors. ISACC alleviates workload burden by automating or supporting tasks from both the clinical and clerical roles. At enrollment time (A), the tool may help with candidate identification and automated message scheduling, and simplify the workflow by automatically pulling in existing patient information if applicable and saving appropriate documentation e.g. the patient's enrollment status back to the EHR. As the program progresses, the tool may automatically send out communications as they become due (B), using the EHR to determine the most current contact information and patient name, and recording what messages were sent and when. The system would also continuously monitor for patient replies (C) and in the event that a participant replies to a scheduled message, prioritize the message for follow-up using NLP, alert staff of the message and its priority, and optionally send an automated response if the implementing organization chooses to do so; during this process, the application will file the message and associated risk score to the EHR. Next, staff may compose a follow-up message (D). ISACC can substantially assist in this process by harmonizing and presenting relevant patient information from several sources. The EHR may contain demographics, diagnoses, depression score history, a suicide safety plan, and more. There may also be intervention-specific information such as intake notes and message history. ISACC may also provide reference information the user may find useful, such as response templates or resources for

Table 8.1 Design opportunities (possible ISACC features)

CC = Caring Contacts

Design consideration	Features (Possible solutions)	Workflow step (Task)
1. Automation Resource scarcity combined with the workload imposed by CC is a barrier. Informatics tools can help by automating simple or repetitive tasks, and using machine learning (e.g. risk modeling, automated information extraction) to help focus human expertise where it is most needed.	candidate identification (eligibility)	A
	automatic scheduling based on template (customizable)	A
	automatic sending of CC based on schedule	B
	automatically triggered follow-ups (e.g. for urgent late night messages)	C
	alerting CC staff of new messages	C
	prioritizing messages for follow-up (e.g. by patient baseline risk, urgency, timing)	C
	automatic routing/delegation (e.g. based on expertise, availability)	C
	automatic extraction of clinically relevant insights from messages	D
	auto-generate clinical note contents based on template	E
2. Information retrieval and synthesis for cognitive support Authoring effective follow-up messages requires consideration of patients' unique circumstances. Different kinds of data from different sources are used for this. Informatics tools can help by aggregating and synthesizing this information, and by making contextualized suggestions.	display current and historical patient info (demographics, questionnaire history, …)	D.a.
	display current and historical intervention info (intake notes, message history)	D.a.
	display current and historical machine learning insights (urgency scores, extracted themes)	D.b.
	display intervention reference information (resources, message templates, …)	D.a.
	suggestions based on patient needs (what resources to include, what messages are likely to be well received, rewrite suggestions)	D.b.
3. Data and workflow integration CC workflows require communication between systems and people with different roles and responsibilities. Informatics tools can help by providing planning tools, facilitating information exchange with external systems, and enabling workflow integration across systems.	Launch workflow from within an existing system	A, B, C, D
	show/integrate data from external systems when appropriate (no double data entry, login into multiple systems)	A, B, C, D
	share patient data back to external systems (e.g. file clinical notes for ongoing care and insurance reimbursement)	A, B, C, D, E
	capture (& share) patient- and intervention-level metrics for reporting	A, B, C, D, E
	patient referral/enrollment status tracking (for handoff)	A
	staff assignments (and temporary delegations)/staff work queue management	C, D
	action logs/task status management	C, D

copy-pasting (e.g. crisis line number). Finally, the current and historical insights extracted from messages with NLP are displayed here, namely risk scores and extracted message themes. The final step (E) is filing appropriate documentation for any action taken. ISACC can help by automatically generating the note text with the patient's enrollment information, the messages sent to and received from the patient, and what follow-up actions were taken within the system. The note may also be filed automatically to the EHR.

8.4 Results

We developed a minimum viable product (MVP) of the Informatics-Supported Administration of Caring Contacts (ISACC) application, implementing the subset of features shown in Supplemental Table 1, using HAPI, SMART-on-FHIR, ReactJS, and Twilio. The code is open source and freely available on GitHub. The following sections contain detailed descriptions of the FHIR representation model, application architecture, and workflow.

8.4.1 FHIR representation model

The requirements for the FHIR data model arise from the application requirements (Table 8.1). Our FHIR data model design follows the FHIR-native approach [23], which involves using FHIR not only for data exchange, but to represent data, state, and business logic. Therefore, the FHIR data model is designed to capture all aspects of the application's data requirements. Figure 8.2 shows these data model requirements and lists the FHIR resources/attributes that can be used for each requirement. The resources reference each other as shown in Figure 8.3. See Supplemental Table 1 for the mapping between application requirements and data model requirements. The following sections describe the usage of each resource in detail.

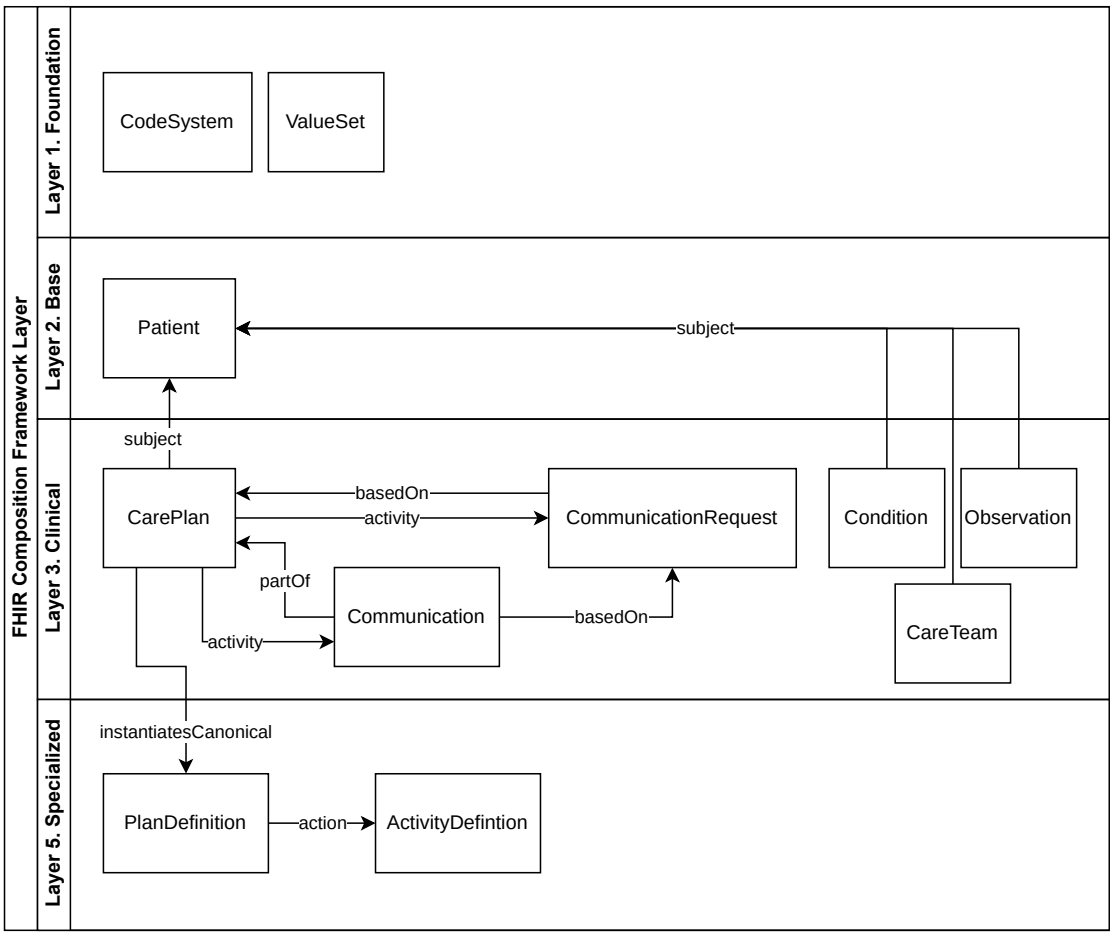

Figure 8.3. Domain model and resource organization. Layer 4 (Financial) is omitted because it is not applicable.

8.4.1.1 PlanDefinition

Caring Contacts involves the use of pre-defined but customizable message plans. For example, a healthcare organization may choose to send a message at one-week intervals, tapering down to two-week intervals after the first month. Many programs additionally send a message on the patient's birthday and on holidays. Similar to the templating approach we developed for StayHome, a COVID-19 symptom tracking app [23], the current work defines a template plan that is instantiated for each patient as appropriate. As described in the Clinical Reasoning Module of the FHIR standard[1], we use the `PlanDefinition` resource to represent the

[1] https://www.hl7.org/fhir/clinicalreasoning-module.html

message schedule template to be instantiated later as a `CarePlan` resource. Each implementing organization develops a custom template according to their goals, preferences, and patient population, and deposits a corresponding `PlanDefinition` resource with the appropriate identifier in the FHIR database. The template is instantiated for each individual patient at enrollment time, allowing customizations of timing and message content on an individual basis. Templated messages are defined in the `action` attribute as `ActivityDefinition` resources.

8.4.1.2 ActivityDefinition

This resource type is used to define individual messages in the default message template, using a value of `CommunicationRequest` for the `kind` attribute. Each message is specified to go out in the specified week of the program, on the specific day of the week, at the specified time of day, using the `timingTiming.repeat` attribute. Further, the message content is defined using the `dynamicValue` attribute, with a `path` value of `payload.contentString` and an `expression.expression` value of the message string, which may contain the "{name}" placeholder, to be replaced with the patient's first name. A limitation of this use of templating language is that the user interface must convey clearly what placeholders are available and what values they will be replaced with; additionally, custom code must be written to perform the substitution.

8.4.1.3 CarePlan

We use the `CarePlan` resource to represent each participant's Caring Contacts plan. This includes the intervention-specific notes (intake notes), which may be amended throughout the course of the intervention but are not episodic, i.e. there are not multiple notes created throughout the intervention; rather, staff maintain one note per patient. However, the note is specific to the current instance of the Caring Contacts program, i.e. if the patient participated at two different points in their life, we would expect to have two different notes, and so the notes should be directly linked to the `CarePlan`. We therefore use the `description` attribute for this

note. The CarePlan resource instantiates the organization-level PlanDefinition, and uses the instantiates attribute to reference this template. The required intent attribute is set to plan. We additionally use the status attribute to represent the patient's enrollment status, utilizing the default value set binding (RequestStatus). The initial value upon creation will be active; program completion will result in a status of complete, and unenrollment without completion will result in a status of revoked. The resource is marked as a Caring Contacts-specific resource using the category attribute. The appropriate patient resource is referenced with the subject attribute. The activity attribute will reference all CommunicationRequest and Communication resources belonging to this particular Caring Contacts plan.

8.4.1.4 CommunicationRequest

Once the message schedule is created, CommunicationRequest resources are created according to the schedule: for each message, the message content and date/time are resolved using the placeholders and patient's program start date. Each resource will contain the message content in the payload.contentString attribute and the planned instance of sending in the occurrenceDateTime attribute. Further, the patient resource is referenced in the recipient attribute. The status attribute is initially set to active, but updated to completed once the request is executed, or to on-hold if the request failed to execute. The CarePlan the resource belongs to is referenced using the basedOn attribute.

8.4.1.5 Communication

Communication resources are created for each delivered message throughout the course of the intervention. When a CommunicationRequest.occurrenceDateTime happens, the messaging API (e.g. Twilio) is invoked to send an SMS message to the phone number with system sms in the linked Patient resource's telecom attribute. Upon successful delivery, a Communication resource is created with a basedOn attribute referencing the

`CommunicationRequest`, a `partOf` attribute referring to the appropriate `CarePlan` resource, and with the `category` attribute marking the resource as a Caring Contacts message. Further, the `payload` attribute contains the message content, the `sent` attribute contains the delivery date and time, and the `recipient` attribute contains a reference to the patient. The `medium` attribute contains indicates that the message was sent via SMS using the `SMSWRIT` code from the `ParticipationMode` (http://terminology.hl7.org/ValueSet/v3-ParticipationMode) code system. Further, the resource status is set to `completed`; the status of the linked `CommunicationRequest` is updated to `completed` as well.

`Communication` resources are also created for incoming messages. When the SMS provider notifies ISACC that a patient responded, the patient is identified via phone number, and a resource is created that references that patient in the `sender` attribute. The `partOf` attribute will reference the patient's currently active ISACC `CarePlan`. The status will be `completed`, the `category` will mark the Communication as a received message, and the `medium` will be set to a `Coding` with system `ParticipationMode` and value `SMSWRIT`. The `sent` attribute will be populated with the current time, and the `payload.contentString` attribute will contain the message text.

The message priority and message themes as extracted by the machine learning modules will be populated in the `priority` and `extension` attributes, respectively. Previous approaches suggest the creation of `Observation` resources for suicide risk scores. However, the risk scores calculated as part of ISACC do not correspond to any standardized idea of risk score, e.g. a score calculated from patient answers on the Columbia Suicidality Screener questionnaire. Rather, they represent a probability of a particular message being urgent, and it may not be appropriate to consider this score outside of the context of that communication. We therefore represent the score using the priority attribute of `Communication` resources, rather than creating `Observation` resources. Future work may explore the use of `Observation` resources if machine

learning can be used to infer suicide risk measurements that are valid and meaningful outside of the message context.

Finally, `Communication` resources are also created when Caring Contacts staff send a manual message to a patient, such as well following up with the patient about a message they sent. After the appropriate messaging API is invoked and the message is successfully delivered, a `Communication` resource is created with a reference to the patient's `CarePlan` in the `partOf` attribute, a `status` of `completed`, a `category` indicating the message type, a `medium` with system `ParticipationMode` and code `SMSWRIT`, the delivery time in the `sent` attribute, a reference to the patient in the `recipient` attribute, and the appropriate `payload.contentString` attribute.

In each case where a `Communication` is created, the `CarePlan` resource is updated to reference it within its `activity` attribute.

8.4.1.6 DocumentReference

We use `DocumentReference` resources to represent progress notes regarding the Caring Contacts intervention. The note text is contained in the `content.attachment.data` attribute as html-formatted text. The format is indicated by the `content.contentType` attribute having a value of `text/html`. The `category` attribute designates the note type as Progress Note (http://loinc.org/11506-3), and also as Caring Contacts note (http://isacc.app/CodeSystem/note-type/isacc-progress-note). Further, the status attribute will be `current`, and the `date` attribute will be the document creation time, and `subject` will reference the relevant patient.

8.4.1.7 Patient

For the purposes of ISACC, `Patient` resources must at minimum contain a `telecom` entry with `ContactPoint.system` value of `sms`, at least one `name` entry with a non-null `given`

attribute, and a `birthDate` which will be used to schedule a birthday message. Further, ISACC will look for emergency contact information in the `contact` attribute, such that the `contact.relationship` has a coding of type `PatientContactRelationship` (http://hl7.org/fhir/ValueSet/patient-contactrelationship) and value `C` (Emergency Contact).

8.4.1.8 Other patient information

ISACC displays the most recent value for `Observation` resources of the following kind, if present: First, those with code 44261-6 (PHQ-9 depression questionnaire scores); second, those with code 93373-9 (Columbia Suicidality Screener score). Further, the diagnosis name (`condition.code.coding.display`), onset date (`condition.onsetDateTime`), and status (`condition.clinicalStatus.coding.code`; e.g. active vs. resolved) are shown for each of the patient's linked `Condition` resources. Finally, name (`participant.member.display`) and role (`participant.role.coding.display`) are displayed for each provider in the patient's `CareTeam`.

Table 8.2 FHIR modeling requirements for ISACC

All data needed to preserve state and support application logic are modeled in FHIR.

Data model need		FHIR resource & attribute	Comments
System-level message schedule template	System-level message schedule template	PlanDefinition(where={id=cc-message-schedule-template}) PlanDefinition.actions Content and time of day for birthday message: PlanDefinition.action(where={ trigger.type=named-event and trigger.name=birthday})	Each message should have message content, time spacing, i.e. how many weeks or months to go between messages, and time of day. Should be able to specify different time of day for each message. Support birthday message.
	Scheduled message	Content: ActivityDefinition.dynamicValue(where={path=payload.contentString}).expression.expression Time spacing & time of day: ActivityDefinition.timingTiming.repeat.period ActivityDefinition.timingTiming.repeat.periodUnit ActivityDefinition.timingTiming.timeOfDay	
Patient/intervention-level message schedule	Patient/intervention-level message schedule	CarePlan(where={intent==plan, status=active, subject=patient, category= http://isacc.app/CodeSystem/careplan-type|isacc-message-plan}) Careplan.activity consisting of CommunicationRequests with: CommunicationRequest.payload CommunicationRequest.occurrenceDateTime CommunicationRequest.recipient	
	Patient/intervention-level notes	CarePlan.description	Intake notes, may be amended later. There can be only one note per Patient per CarePlan
	Enrollment status	CarePlan.status active completed	status==revoked if program was not completed (e.g. if patient wishes to no longer receive messages)

		revoked	status==complete if all messages in message schedule have been sent
Message record	Message priority	Communication.priority routine urgent	urgency level (priority) for each received message
	Message response status	Communication(where={recipient=patient, sort=-sent, _count=1}).sender	
	Message response time	Communication(where={recipient=patient, sort=-sent, _count=1}).sent	To find the amount of time since patient-initiated message was received
	Message type	Communication.category http://isacc.app/CodeSystem/communication-type | isacc-auto-sent-message http://isacc.app/CodeSystem/communication-type | isacc-manually-sent-message http://isacc.app/CodeSystem/communication-type | isacc-received-message	Distinguish automatically sent messages from manual messages
	Message delivery status	Communication.status completed if successful not-done with Communication.statusReason==system-error if there was a technical failure preventing successful delivery	Keep track of message sending success for outgoing messages
Progress note record	Note type, date, and content	DocumentReference(where={category=http://loinc.org|11506-3 or category= httpn://isacc.app/CodeSystem/note-type | isacc-progress-note}) DocumentReference.date DocumentReference.attachment.data	File patient note to EHR when response authoring workflow is being finished. Mark the note as a Progress Note (LOINC 11506-3) as well as a caring contacts note for easy retrieval.
Patient record	Patient name	Patient.name.given	Name to be used to address patient in message. The patient may in theory have a preferred name that is different from their given name. The use attribute requires NameUse and does not allow extension, but a future version may consider adding a Caring Contacts -specific name preference, for example as an extension.

	Patient contact information	Patient.telecom(where={use=sms})	Contact details to which Caring Contacts messages should be delivered (channel e.g. SMS, phone call; actual contact detail e.g. phone number, email address). Only SMS for now; in the future, add use=phone/use=email. Patients may have multiple contact details, even multiple phone numbers. Some phone numbers may not be private (e.g. a shared phone) or may not have the ability to receive SMS. Need the ability to designate a contact point for Caring Contacts that may be different from the contact point that is preferred for other doctor's office communication.
	Patient birthday	Patient.birthDate	To schedule birthday message, and for identification purposes.
Observation records	Observation name, date, and value	Observation.code Observation.value Observation.effectiveDateTime	
Diagnosis records	Diagnosis name, date, and status	Condition.code.coding.display Condition.onsertDateTime Condition.clinicalStatus.coding	
Care team records	Name, role, and contact information for each care team member	CareTeam.participant.member CareTeam.participant.role.coding CareTeam.participant.telecom	

8.4.2 Application architecture

The system components are shown in Figure 8.4.

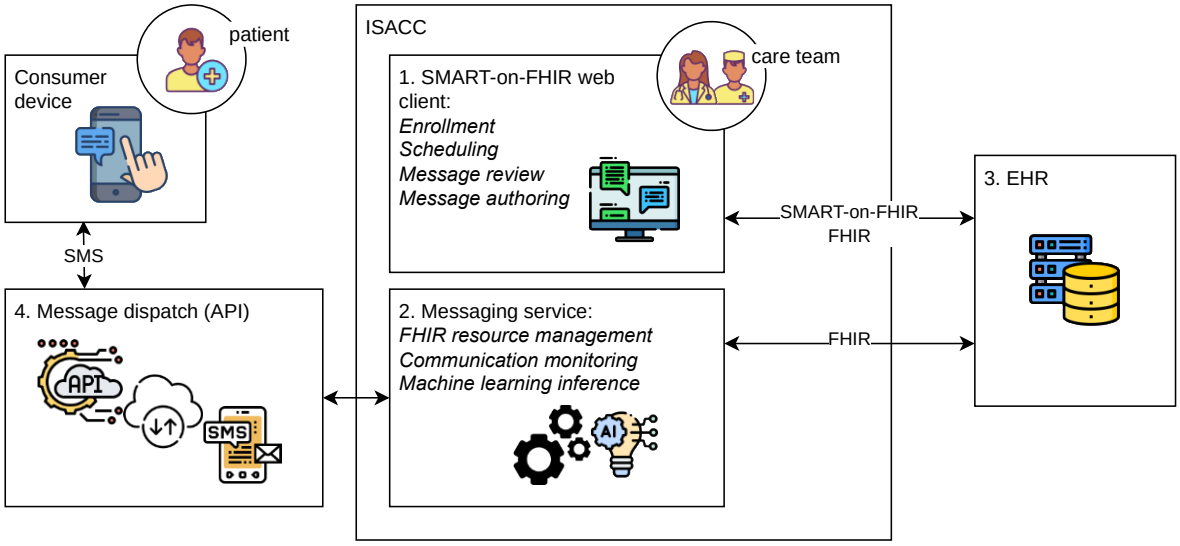

Figure 8.4. ISACC application overview.

The system involves 1. a SMART-on-FHIR web client application; 2. a server app asynchronously providing the message service for sending and receiving messages as well as running machine learning models; 3. a host system implementing FHIR APIs, such as an EHR; 4. a message dispatch service, such as Twilio, for sending and receiving messages. Patients receive messages as SMS text messages at their designated phone numbers.

8.4.2.1 SMART-on-FHIR web client

This is the user-facing application providing functionality for enrolling, managing, and communication with patients in the Caring Contacts intervention. The user begins by initiating the application launch from within the FHIR host system, such as an EHR. This establishes an appropriately authenticated connection between the ISACC application and the EHR, allowing the application to collect patient information from the relevant EHR chart, including first name, last name, and date of birth for identification purposes. Contact details, as required by ISACC to send messages, as well as emergency contact information are also retrieved from the host system, or can be entered if not available.

At enrollment time, a message schedule is generated from a configurable template defined on the system level. Users can then remove or add messages and change the messages'

text and date/time, as needed. The patient's first name can be included within the message text using a placeholder. Users can also enter patient- or program-specific notes, such as intake notes. Once the user completes this workflow, the schedule is created in the host system's database as a CarePlan via FHIR API, and CommunicationRequest resources are generated accordingly.

Once the patient has been enrolled and their messages scheduled, users can access the messaging view. This view includes patient information for reference: that entered during enrollment, as well as any data available in the linked EHR, which may include the patient's suicide safety plan, PHQ-9 (depression questionnaire) scores, Columbia Screener (suicidality) scores, diagnosis information, and details of the care team. The application also provides a view of all relevant messages (Communication resources) that have been exchanged so far. Using this view, the user can also send messages to the patient, such as responses to patient replies or one-off communications.

8.4.2.2 Messaging service (server app)

This continuously running server app manages FHIR resources, dispatches and receives SMS communications, and utilizes available machine learning models to draw and store inferences about the messages. To accomplish this, it is in continuous communication with the host system via FHIR APIs, polling for `CommunicationRequest` resources and executing requests as appropriate. When `CommunicationRequest` resources are triggered, the service creates the message, obtains the appropriate contact details needed to send the message, and requests the message be sent via an API to the communication platform. Once the SMS API reports that a message was sent, the service creates a `Communication` resource with the details and status of the message and marks the `CommunicationRequest` as complete. The server app also acts as contact point for the SMS platform, which passes along incoming SMS messages via a call to a defined webhook. Once such a message is received, the service uses a machine

learning model to draw inferences (calculate a risk score), then creates a `Communication` resource containing the message details and inferences. The `Communication` resource is deposited in the FHIR database and will be thus be shown in the web client app when it is next viewed.

ISACC uses the best-performing BERT model developed in Chapter 6. The fully trained model is deposited in the file system and thus available to the application at runtime. Upon receipt, the model is loaded into memory and used to predict a binary urgency label for the received message. A label of 1 is represented as a priority of `stat` in the Communication resource; `routine` is used for a label of 0.

Of note is that the SMART-on-FHIR web client and the messaging service app do not stand in communication with each other. While each application interacts with the FHIR database, they are independent actors in the system. This allows the SMART-on-FHIR web client to maintain its status as confidential client, i.e. an application that only interacts with the FHIR host and no other outside service.

8.4.2.3 SMART-on-FHIR host (EHR)

This could be an EHR system such as Epic or Cerner. However, a significant advantage of using FHIR and SMART-on-FHIR, which are standardized protocols, is that ISACC is completely portable between systems as long as FHIR and SMART-on-FHIR APIs are provided. If an EHR system is not available or if the available system does not implement all the required APIs, a host system harness can be used. ISACC was developed using fEMR, our open-source, "free-standing" SMART-on-FHIR host system internally developed for another project[1], which can be used for development and demo purposes; fEMR could also be used as the primary host system in an operational implementation of ISACC. While this reduces the benefits derived from

[1] https://github.com/uwcirg/cosri-environments

using FHIR and SMART-on-FHIR – for example, separate credentials would have to be maintained and duplicated data entry efforts may be required – it represents a useful waypoint along the pathway of EHR integration.

8.4.2.4 Message dispatch system

Here, we demonstrate the feasibility of ISACC using SMS for communication using the commercial messaging vendor, Twilio[1]. However, another API-based communication platform could be substituted. For example, a system that sends and receives communications via email instead of SMS. A traditional postal mail process for Caring Contacts could be incorporated as well, as long there is a way to dispatch outgoing messages (e.g. a task to printing service or work-queue for a post card to be written by hand) and optionally, receive incoming messages.

8.4.3 Workflow overview

This section illustrates the user-facing functionality that ISACC implements to support the workflow as shown in Figure 8.2. There are three views: the patient list, the patient enrollment view, and the messaging view.

8.4.3.1 Patient list

The patient list view (Figure 8.5) is the launch point for the Smart-on-FHIR application, and may thus be part of the EHR system that serves as application host. In standalone systems, this FHIR host system role is fulfilled by our SMART-on-FHIR harness app, called fEMR. This view shows a list of all patient records in the system. In fEMR, new client records can be created by entering first name, last name, and date of birth. Once in the system, patients can be enrolled in a program (that is, a message schedule can be created); to begin this process, users launch the SMART-on-FHIR web client app in enrollment mode from the patient list view. To review

[1] https://www.twilio.com/

patient messages and to send new messages, users launch the SMART-on-FHIR web client app in messaging mode. In fEMR, the patient list view also includes a column displaying if a response has been received but not reviewed, and if so, how long ago that messages was received. If the message was received more than 1 business day ago, this column also indicates this with an additional flag. Another column shows an indicator icon if the outstanding message appears to be urgent. The urgency and time indicators will only be shown until Caring Contacts staff follow up. Deployment of ISACC within a commercial EHR system will require corresponding column configuration according to the EHR vendor's instructions.

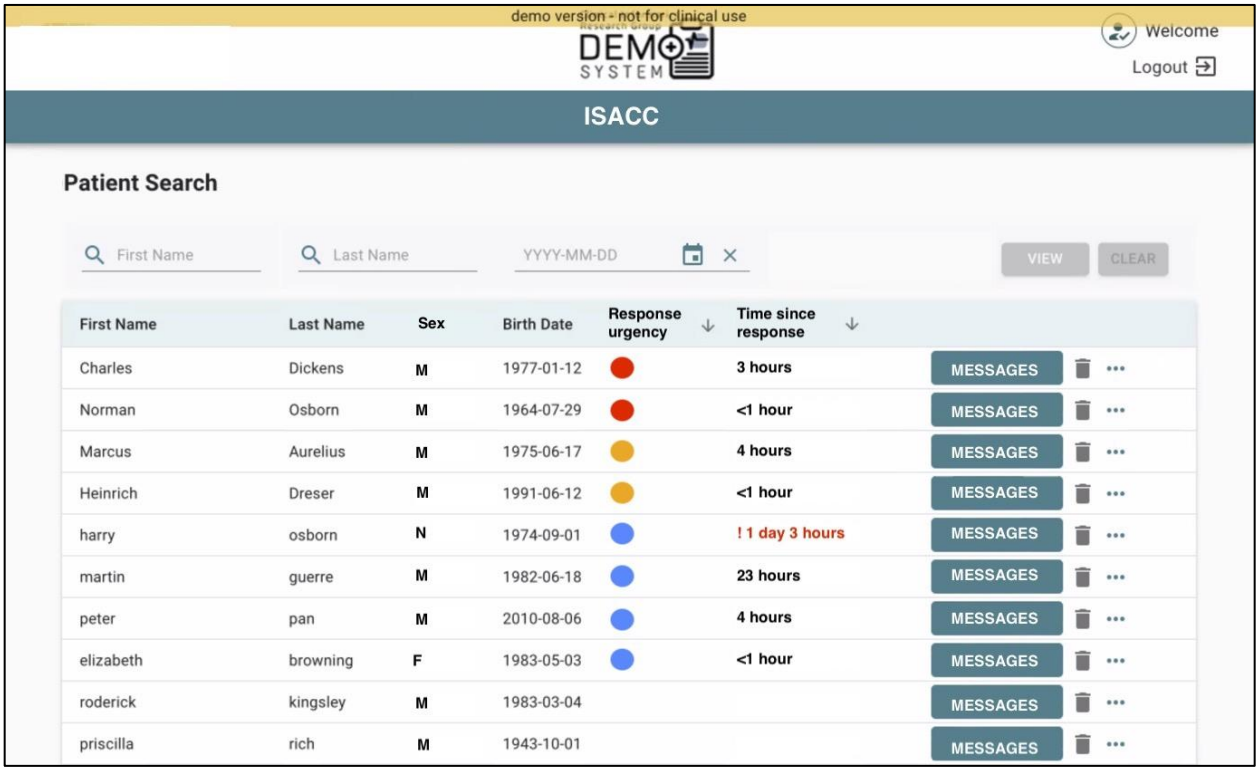

Figure 8.5. Patient list

8.4.3.2 Patient enrollment

The enrollment view (Figure 8.6) allows the user to specify the message schedule for a patient being newly enrolled in the intervention. Initially, the default message schedule (configured by the organization in a system-level template) is shown, which contains the default timing and text for each message. Date, time of day, and message content can then be

customized for each message, if desired; messages can also be added or removed. Users may use placeholders, such as “{name}” to substitute the patient’s preferred name. A birthday message is generated as part of the schedule if the patient's date of birth falls between the first and last scheduled message. Patient intake notes can be entered here as well. Completing this enrollment step will cause appropriate records to be created in the database, which will then trigger SMS messages to be sent as specified.

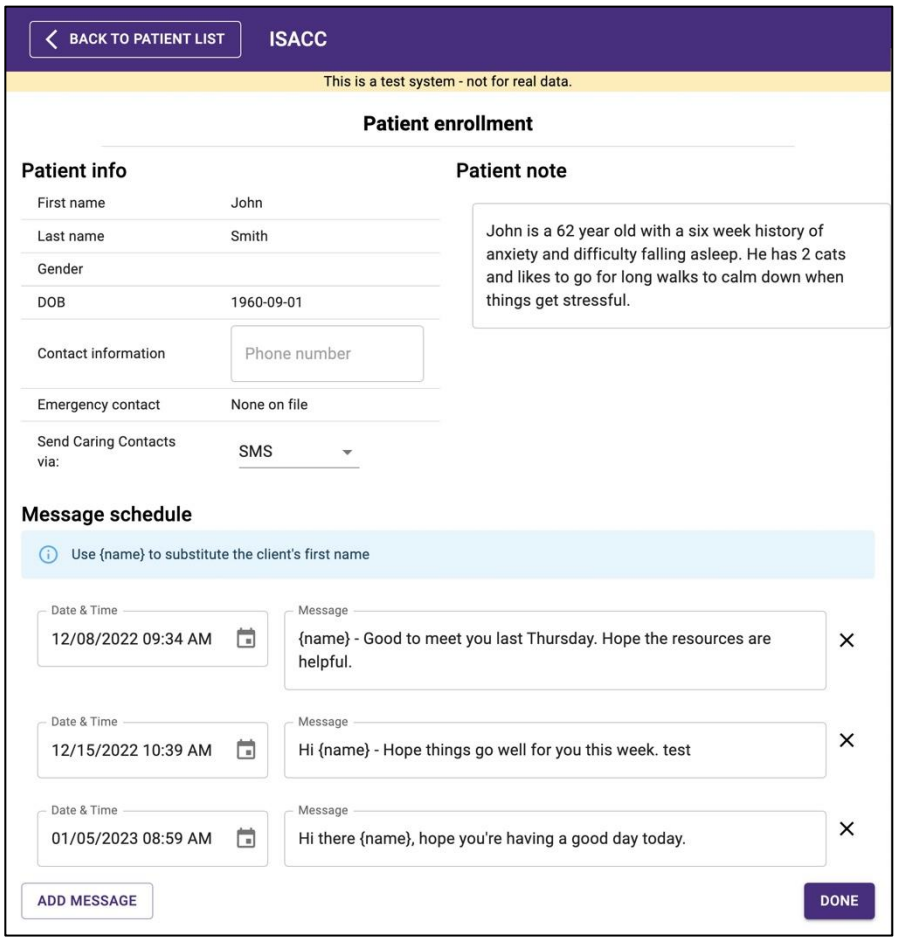

Figure 8.6. Enrollment view

8.4.3.3 Messaging

The message view (Figure 8.7) shows the history of exchanged messages and allows users to write replies. Patient information and patient notes are shown for reference, and can be amended if needed. If available, e.g. from a linked Electronic Health Record system such as

Epic, the most recent PHQ-9 (depression) and Columbia (suicidality) questionnaire scores, diagnoses, and provider information are shown. Messages that the system determined to be urgent are highlighted. Language use patterns and expressions that may be related to suicide risk (e.g. entrapment, hopelessness) are flagged. Users can write new messages, which will be shown in a different color to differentiate them from the scheduled, automatically sent Caring Contacts messages.

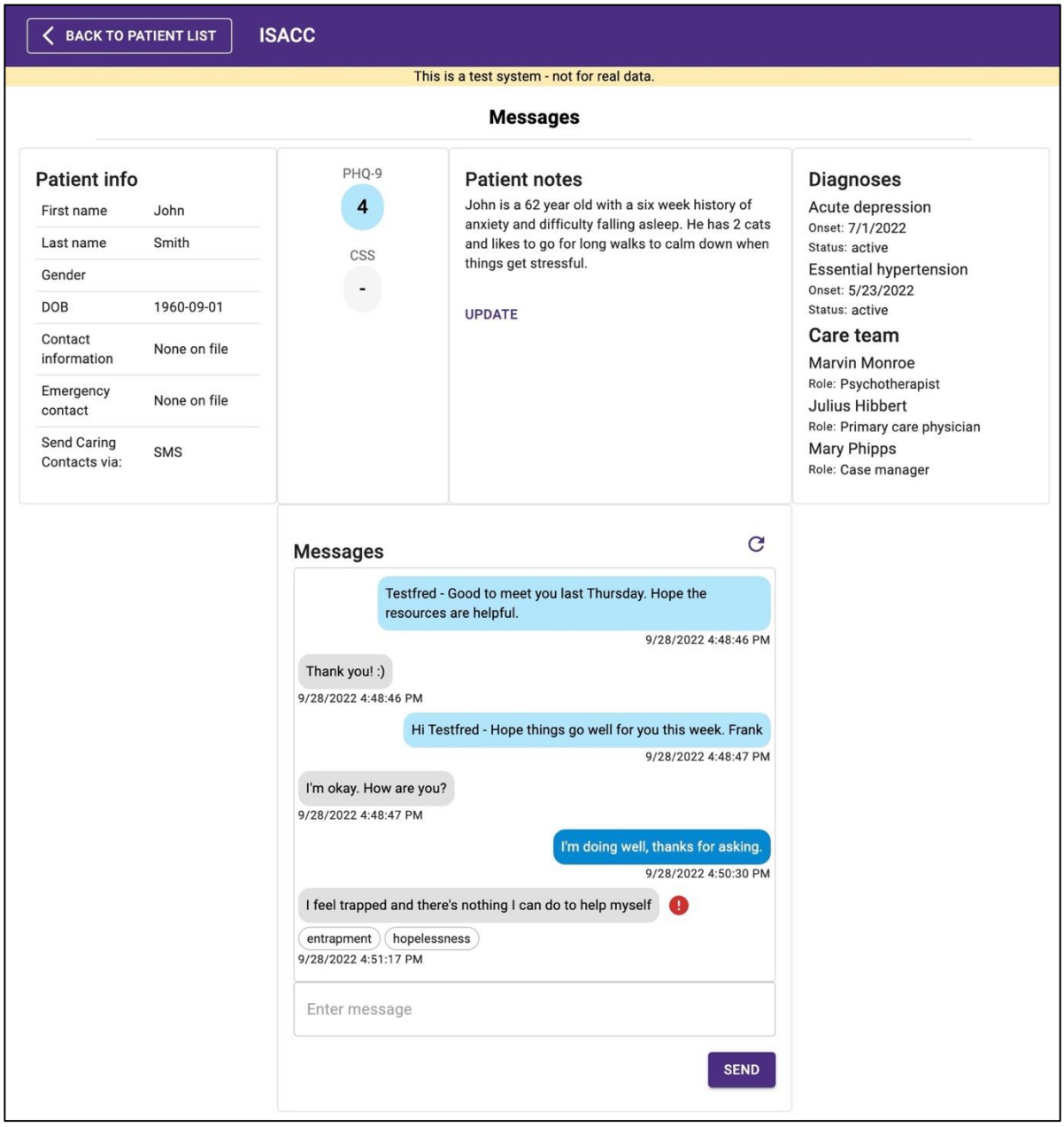

Figure 8.7. Messaging view

8.5 Discussion

In this work, we used findings from a previously published needs assessment study [221] to develop a list of possible features that would address these user needs. We then developed application requirements for an initial implementation of such an application, selecting a subset of features to implement for the purposes of a minimum viable product (MVP) for ISACC. Based on these functional requirements, we determined the data modeling needs and assessed the feasibility of using FHIR to meet these needs, demonstrating that FHIR is capable of accounting for all the data model needs for Caring Contacts arising from this and prior work. We additionally developed an architecture for an information technology system capable of ingesting PGD and making machine learning inferences in real-time, honing SMART-on-FHIR to enable portability to commercially available EHR systems. Finally, we implemented a functional MVP embodying these designs, demonstrating the feasibility of the system. We provide the MVP in a freely available, open-source repository under the BSD 3-Clause license.

8.5.1.1 A FHIR architecture for PGD and machine learning

To the best of our knowledge, there have been no reports of applications that automatically ingest and process PGD and integrate them into clinical workflows in a meaningful way. This is because of challenges ranging from system architecture questions to data representation and workflow integration questions. SMART-on-FHIR is promising, but there are unresolved questions of how it can be leveraged to enable workflow integration in this context, and how it might be set up to continuously query for external data. Using FHIR as basis for data modeling raises questions of data representation choices, and questions of where and how this data is deposited in a FHIR database. Finally, there are workflow integration considerations that will directly affect clinical utility, i.e. of how external data can be made not only clinically meaningful, but also actionable within clinical workflows..

This work represents a proof of concept for this use case. We designed an architecture for a SMART-on-FHIR-based application system that includes a continuously running, asynchronous service that stands in communication with the source of PGD via API, processes incoming data using machine learning in order to add clinical meaning, and translates it into FHIR resources. We further designed and developed a user-facing client application according to findings from a systematic user needs assessment, integrating PGD and insights generated from it into an existing workflow that stands to benefit from this. This system serves as a blueprint for similar systems that may use different kinds of PGD, applying different kinds of AI, and solve different clinical problems.

8.5.1.2 ISACC Requirements

The previously published needs assessment determined that any informatics tool supporting the Caring Contacts suicide prevention intervention should focus on three core needs: full or partial automation of labor-intensive, time-consuming tasks; providing cognitive support to human experts as they complete clinical tasks; and providing data and workflow integration to facilitate team communication and reduce the workload burden imposed by workflows being spread across multiple information systems. ISACC addresses these needs in the following ways.

1. Automation
    a. ISACC allows the definition of a message schedule template on the organization level, with message content and timing for each message. The template is instantiated at enrollment time, using the patient name, birthday, and enrollment date as appropriate. The creation of a message schedule is therefore simplified because neither message text nor dates and times have to entered manually. The

process of setting up the message schedule for a newly enrolled patients is thus reduced to only a few clicks.

b. ISACC automatically sends out messages according to the schedule via SMS. Staff members therefore do not have to manually keep track of schedules and send messages by hand on the planned dates, removing the daily task of sending out scheduled messages entirely.
c. ISACC alerts designated staff if a patient replies to a Caring Contacts message, removing the need to check for replies at regular intervals.
d. ISACC automatically triages messages for follow-up, allowing staff to be more flexible with time spent on triage and composing follow-up messages. In contrast to current workflows, staff will only have to review a small subset of incoming messages in an urgent manner if ISACC is used, and can follow up with patients who do not need an urgent response at a later, perhaps more convenient time, e.g. in batch.
e. ISACC allows users to auto-generate progress notes for the purpose of clinical documentation. In contrast to current workflows, where staff have to type up clinical notes manually, this task is reduced to only a few clicks.

2. Cognitive support
    a. ISACC displays selected patient information for reference. In contrast to current workflows, where staff may have to review large parts of patient charts, this supports the cognitive processes involved with clinical reasoning and decision making by removing the cognitive burden of mentally filtering information.
    b. ISACC allows staff to review and amend intervention notes in the same view where message history is reviewed and follow-up messages are composed.
    c. ISACC highlights messages that appear to indicate elevated risk. This directs clinicians' attention, improving efficiency of the message review process.

3. Data and workflow integration
    a. ISACC is launched directly from within the patient context in the EHR, removing the need to log in to a separate system and search for the correct patient record. In contrast to information technology ecosystems with disjointed platforms, this improves workflow efficiency and removes opportunities for errors, e.g. opening the wrong patient record.
    b. ISACC uses a single, collaboratively created intervention note per patient, which serves as a place to communicate about the intervention strategy, e.g. things to keep in mind about the patient.
    c. ISACC retrieves all patient data, including text messages and questionnaire scores, from linked FHIR databases, e.g. patient demographics from the EHR. This removes the need for duplicated data entry of the patient's emergency contact information, primary care provider name and contact information, and other relevant information that is already present in the EHR.
    d. ISACC uses the FHIR host's database system as primary data store, i.e. data created by ISACC is automatically deposited into the EHR. As a result, records created by ISACC, e.g. CarePlan and Communication resources, are available to and findable by the host system (e.g. EHR) and other applications utilizing the EHR's database. For example, depending on EHR configuration, the Caring Contacts plan may automatically appear in the EHR's CarePlan tab, and messages may automatically appear alongside other patient communications in the appropriate sections of the patient chart in the EHR. Similarly, message urgency scores are saved in this database. In contrast to current workflows, this allows other care team members, such as behavioral health providers, to easily review intervention progress.

e. Because ISACC uses the FHIR host's database system as its primary database, all data created and used by ISACC is available to reporting tools operating on this database, e.g. the wide range of reporting tools that are available as part of commercial EHR systems. For example, program coordinators could use the reporting tools they are already familiar with to write reports for the number of messages sent and received, or the amount and content of urgent messages and the timing of responses to such messages.
f. Because ISACC uses the host system's patient list functionality, patient list functions widely available as part of commercial EHR systems can be easily leveraged to benefit the Caring Contacts workflow. For example, users may be able to create patient lists that only contain the patients they are responsible for, or they may use extensive column configuration options to customize the patient list display according to their needs.

8.5.1.3 Alternative designs

This section describes some possible alternative approaches for the design decisions that we made and provides further reasoning for our design choices in light of these alternatives.

8.5.1.3.1 FHIR representation choices

In this work, we used the FHIR-native approach [23], i.e. we used FHIR as our primary and only data store. In this approach, FHIR is used to model every aspect of the application. We demonstrated that this approach is feasible for ISACC. Thus, we realized the benefits of the FHIR-native approach, e.g. removing the need for a separate database. However, the design choices enabling this have tradeoffs. FHIR is flexible and comprehensive, but it is intended for clinical data, which limits what data may be appropriate for FHIR. For example, the use of the `Communication.priority` attribute to store message urgency is reasonable and appropriate, as urgency informs the order in which messages should be read and addressed by users, similar to

e-mail priority flags. However, `priority` requires the use of the appropriate value binding, which limits the possible values to `routine`, `urgent`, `asap`, and `stat`. Our machine learning model is a binary classifier, i.e. it outputs a probability between 0 and 1 for each of two possible classes, and we chose to store the most likely class prediction using one of two priorities: `routine` or `urgent`. This is appropriate, as these two values fully capture the level of detail needed within the clinical workflow. We do not store the raw probability calculated by the model. However, while this raw number may not be immediately relevant in a clinical context, it is possible to imagine a use for this score, e.g. for error analysis of the model's real-world performance. Therefore, by reducing the data to elements deemed clinically relevant as per the FHIR specification, there is information loss. We could counteract this loss by utilizing a separate, ISACC-specific database to be used only for data that is not appropriate for the clinical data store; or, we could store the raw number as a generic resource extension. However, in keeping with the parsimonious intentions of FHIR and the FHIR-native paradigm, we chose not to do so.

The urgency is a proxy for suicidality, so one alternative to using `Communication.priority` might be to store it as a "suicidality score" in an `Observation`. However, we felt that the calculated risk/urgency score is not sufficiently meaningful outside of the `Communication` context to justify an `Observation`. Although it is valid to define an internal code set for the application, `Observation` resources are intended for standardized observations (i.e. those of a type belonging to a controlled vocabulary and, therefore, a specified clinical). It is valid to define a local vocabulary for ISACC, but this might limit the discoverability or usefulness of these resources outside of ISACC, e.g. in the EHR context. Another alternative is the `Condition` resource, representing a patient's suicidal ideation as a condition akin to a diagnosis, problem list item, or symptom. Conditions have onset (start) and resolution (end) dates, which may be convenient as the beginning and end of a period of time where a patient requires specialized support, as defined per clinical protocol. However, it is not trivial to infer patient

state, including start and end times of that state, from individual patient messages. For example, if a patient were to send a concerning message, but then immediately follow up with a positive message such as “Never mind, I’m okay” or “But otherwise I’m doing really well”, one might argue that the `Condition` be marked as resolved, but it may be equally valid to keep it active. Such an inference of the clinical implications of an urgent message may not be valid, or may not be in line with the implementing organization’s values. Representing the algorithm’s outputs strictly as what they are, i.e. predicted urgency of an individual incoming text message, avoids making possibly invalid assumptions about the value’s implications. Another alternative is the `Task` resource, which represents a generic workflow step and its status. Here, it may be used to represent the follow-up task, i.e. a `Task` resource may be created to represent the need to follow up with a patient who may be in distress, and marked complete once follow-up is completed. This would be a direct representation of the workflow implication of an urgent message, which is only implied by representing the urgency score itself. However, this approach is not robust to variations in organization-specific intervention design decisions. For example, if organizations decided that not all urgent messages should trigger follow-up, or that non-urgent messages should be followed up as well, then this representation model would have to be adjusted.

Many clinical data elements may be present in different types of observations, and it may not be straightforward to find the specific information one is looking for. Such is the case here with `CareTeam`. There are numerous possible approaches to finding practitioners that are associated with a given patient, e.g. by finding all providers linked to the patient’s encounters. However, this would yield many practitioners that would not be good contact points for the patient in the event of a crisis, e.g. those with a limited scope of practice such as physical therapists, or those who are not part of the patient’s local care team, e.g. a doctor they saw when they got sick on vacation that one time. Another alternative is using patient attributes such as `generalPractitioner`. However, this attribute would not yield all the people that might be appropriate in the context of suicide prevention, which includes social workers, case managers,

and behavioral health providers. Similarly, filtering providers by specialty can be tricky because there is a very long if not infinite list of specialties, depending on the value sets used, and it may not be possible to foresee how they are used by a particular organization. Ultimately, the best ways to query for a given patient's providers will depend on how data is represented. Here, we chose to use the `CareTeam` resource, which is a curated resource of practitioners who are responsible for a given patient in a given context. The main limitation of this is that an appropriate `CareTeam` resource may not exist, or the resource may not include the appropriate list of practitioners. However, with this approach, there is a possibility of allowing ISACC users to easily edit the resource to be fit for use in the Caring Contacts context.

8.5.1.3.2 Using SMART-on-FHIR

A traditional approach may have conceptualized this system as a standalone system with its own credentialing system and database. This approach has advantages. For example, it would not depend on any host system being available and providing the appropriate APIs. The system's database would be independent, allowing a higher degree of control over database contents and structure. Minimal data representations could be used that do not contain extraneous elements that may not be applicable to the Caring Contacts use case. However, there are significant drawbacks to this design that severely limit its clinical usefulness. Because credentialing would be separate, users would have to maintain usernames and passwords for multiple systems. Without a launch point within the EHR, users who had been reviewing patient information within the EHR would have to search for and open the corresponding patient record in the separate system. This is not only effortful, but also carries a risk of opening the wrong patient's record. The existence of separate databases would necessitate managing patient information in two different places, and would require users to manually keep shared elements in sync and establishing a process for handling conflicts. Further, it would be difficult to share intervention information, e.g. enrollment status, message content, or calculated risk scores, with the patient's

healthcare providers, as this would involve import/export processes or manual transcription. This approach significantly reduces ancillary data management tasks and workflow overhead, and therefore supports intervention scaling.

8.5.1.3.3 Using patient portals to support Caring Contacts

Another approach to designing a Caring Contacts system is to set up automated patient portal messages. This approach has the advantage that it can be achieved without introducing additional technology components. The entire setup can be achieved via EHR configuration. However, EHR configuration can be time consuming and costly in itself, as it requires specialized expertise and often requires the involvement of vendor representatives. Additionally, the patient experience is changed if messages arrive within a patient portal. In order to see the messages, patients would have to first log in to their patient portal app on their smartphone or on a computer, which can be a barrier if internet access is limited, or if patients expect that the messages are routine and "not important". This would compromise intervention effectiveness; a key component is to be undemanding, i.e. to minimize the effort required on the patient's part to partake in the intervention, and leveraging communication channels that they are used to and already use anyway is a key component of that.

8.5.1.3.4 Using "simple" database solutions

Another possible alternative is to operate a simple database, e.g. as a spreadsheet or Access database, to keep track of individuals and their messages. While this approach does not introduce any additional technology and therefore minimizes technology costs, it is associated with significant workload burdens. Spreadsheets would have to be manually updated, as messages are sent via separate pathways. Searching for and sorting patients would be simple enough, but more advanced functionality might be difficult or impossible. For example, data validation to prevent duplicated or invalid entries, e.g. due to typographical errors, may be possible only with additional technical setup and configuration, and may require users to have

certain technical skills. While this approach may be sufficient if only a few patients are enrolled or if users are well-trained, it is unlikely to scale well.

There is also no opportunity to impose a workflow on such a database. The power of custom applications is that operating procedures and workflows can be encoded, reducing the need for users to manually keep track of tasks and remember the steps to performing these tasks. Additionally, advanced functionality such as machine learning components may be difficult or impossible to incorporate in such solutions. Shareable intervention-specific solutions not only make this possible, but facilitate broader intervention adoption by providing out-of-the-box solutions to common problems, as described in the next section.

8.5.1.4 Clinical utility

Workflow integration considerations are a major factor in developing tools that effectively support clinical workflows and decision making. Recommendations made by guidance such as the "Ten Commandments for Effective Clinical Decision Support" [162] and the "Five Rights of Clinical Decision Support" [252] include avoiding introducing additional workflow steps e.g. by requiring users to switch to a different context, and providing information at the right time within the workflow (where it is neither too early nor too late to act on it), via the right channel (presentation mode, platform, etc.), to the person who actually has the ability and capacity to act on it. These and other issues apply to AI-based CDS. Safety, actionability, and utility have been identified as core issues that must be addressed to achieve adoption of AI into healthcare [14]. Workflow issues may preclude actionability of AI outputs, such as when inferences are not presented at the right time or in the right context, or if outputs are not connected to interventions that may modify them. Poor workflow integration may also affect utility, such as when clinicians disregard AI outputs because they are presented too frequently or intrusively (i.e. alert fatigue), or because obtaining AI-based insights requires clinicians to take extra steps.

Raw PGD are difficult to integrate into clinical workflows in a meaningful way. They often come in large volumes, e.g. in the case of sensor data from wearables, and it is often not immediately apparent if they contain clinically relevant information. As a result, reviewing these data may not have a favorable cost-benefit tradeoff when time is limited. This constitutes a significant limitation on how useful such data can be for the purpose of supporting clinical care via CDS. While summary statistics and data visualization may help clinicians determine the relevance of numeric data quickly, this is more difficult with natural language data. Patient-generated natural language data, e.g. social media data or patient-clinician communications, must be processed using NLP, lest they must be read and assessed individually.

Processing PGD to distill clinically relevant, actionable insights, and appropriately integrating these insights into clinical workflows, is thus paramount to harnessing their potential for CDS. However, these are difficult problems, requiring careful human-centered design and model development efforts. Furthermore, the variability in health information technology infrastructure and operating procedures between organizations might necessitate that workflow integration is designed and developed anew for each site.

Developing reusable, interoperable CDS systems that define how clinically relevant insights can be extracted from raw PGD and integrated into a workflow, in a manner that is robust to variations in technology infrastructure, can reduce the effort required to develop effective integration. Although solutions still have to be tailored to each individual site – for example, models may have to be retrained to perform well on certain patient populations, or workflow parameters may have to be modified to fit an organization's unique vision – shared solutions can provide a starting point that improves the feasibility of effective AI-based CDS by reducing design and development costs.

8.5.2 Limitations

This work should be considered in light of its limitations. SMART-on-FHIR applications are intended to work in conjunction with EHR systems. While this is a significant strength, as it has the potential to address a wide range of barriers encountered in the realm of health information technology, it requires that both parties fulfill their roles appropriately. As of 2020, vendors are legally required to support FHIR APIs per the 21st Century Cures Act [145] for the US Core data elements [260], but current implementations have many limitations. For example, few EHRs give write access to external applications. This is expected and perhaps appropriate, as allowing external applications to modify the contents of databases used for ongoing clinical care raises concerns of data governance and quality. Healthcare organizations would have to ensure that data originating with external applications is correct and complete. EHR vendors and healthcare organizations must define their own business rules for validating resources before allowing them to be deposited into their databases, and this process is time and resource intensive. Therefore, there are currently no clear expectations of the timelines for enabling write access in prevailing EHR systems.

However, this limitation should not discourage the wide adoption of FHIR and SMART-on-FHIR. Designing and developing tools according to accepted standards ensures that tools will work without the need for extensive software changes once the appropriate APIs are made available by vendors. In the meantime, the portability afforded by FHIR, i.e. the fact that tools can work with a wide range of different host systems, enables developers to provide their own FHIR host system harnesses, e.g. our fEMR, with the expectation of supplanting them with an operational EHR system when possible. Another possible transitional solution is to integrate with an EHR system to read existing data, but to use a separate FHIR server for maintaining application-specific data. Because the FHIR interface is identical across FHIR-compliant database servers, such a configuration is readily achievable by choosing the appropriate server address for each database transaction.

8.5.3 Future work

This works enables further research investigating the use of PGD in FHIR-based CDS tools. Specifically for text message-based interventions, future work may directly build upon the FHIR representation model and application architecture we presented, reusing the code that we made available.

In the context of Caring Contacts, this work informs the design and development of informatics tools to support this intervention. By providing the artifacts of our design and development process, we enable healthcare organizations with diverse needs and regulatory constraints to build on our work to develop tools that meet their needs. They may use our design considerations, software specification/suggested feature list, they may use the source code for our implementation as a starting point to write their own application, or they may use our application as is.

Future work may also explore further additions to ISACC functionality. We implemented a minimum set of features in this iteration, but there are numerous further opportunities to support users, including two promising applications for AI: extracting clinically actionable message themes, and writing support for follow-up message composition. Clinically actionable message themes may include linguistic markers of depression or suicidality that may serve as intervention targets. For example, in prior work [135] we developed an approach to extracting metrics of behavioral activation, i.e. the degree to which individuals take an active part in their own lives by planning and participating in pleasurable activities, and validated their correlations with symptom severity in a large sample of written messages exchanged between patients and psychotherapists in teletherapy; ISACC could make use of such techniques to alert intervention staff to an opportunity to encourage patients to participate in activities they used to enjoy. Writing support may be provided in the form of automatic message scoring for being "in the spirit of Caring Contacts" i.e. being empathetic but undemanding; recent work by Sharma et al. [139,140] demonstrated not only the ability to measure empathy from natural language, but to

suggest changes to a given message that would make it more empathetic. ISACC could make use of this to assist intervention staff in writing high-quality follow-up messages.

For the ISACC project, a pilot test is our next step. To prepare the ISACC application for deployment, we will further refine it using scenario-based usability testing, with scenarios based on the specific intervention designs that our participating pilot clinics plan on implementing.

8.6 Conclusion

Suicide prevention interventions face many unique challenges that may be addressed with carefully designed digital tools. Designing and developing such tools with interoperability and reusability in mind is critical to facilitate broad adoption and maximize the benefits for suicide prevention. Also, while patient-generated natural language data represent an opportunity in the Caring Contacts intervention, much work remains in investigating how to effectively integrate them into clinical workflows for the purpose of NLP-based clinical decision support. We developed blueprints for a FHIR-native data representation model and a SMART-on-FHIR application architecture for a software system supporting the Caring Contacts suicide prevention intervention, leveraging patient-generated data and machine learning to support clinical workflows, and demonstrated their feasibility by using them to implement and share a functional open-source application. The data model, application architecture, and implemented tool exemplify approaches to facilitating adoption of AI in healthcare and contribute toward broader adoption of Caring Contacts.

8.7 Acknowledgments

This work was supported by Innovation Grant “Informatics-Supported Authorship for Caring Contacts (ISACC)” from the Garvey Institute for Brain Health Solutions.

Figures in this work have been designed using resources from Flaticon.com.

In the preceding chapters, I developed an informatics tool for Caring Contacts, a text message-based suicide prevention intervention, focusing on assessing workflow and AI needs comprehensively, evaluating clinical utility early on, and leveraging health data standards where possible. The approaches I adopted aim to address recurring challenges facing AI applications for healthcare, and are generally applicable to a range of use cases for AI-based informatics tools. The next section therefore describes a generalizable framework that incorporates recommendations for a needs-driven, utility-oriented, standards-based approach to operationalizing AI in healthcare.

Chapter 9. A framework for needs-driven, utility-oriented, standards-based AI in healthcare

In this work, I developed and applied a generalizable framework for the needs-driven operationalization of artificial intelligence (AI) to support healthcare workflows and clinical decision making. This comprehensive end-to-end framework is intended to guide biomedical informatics projects involving AI. An extension to a framework by Jung and colleagues [15], it spans the design, development, and implementation of AI for clinical decision support (CDS). I propose three specific extensions to the framework, shown in the boxes labeled 1-3 in Figure 9.1.

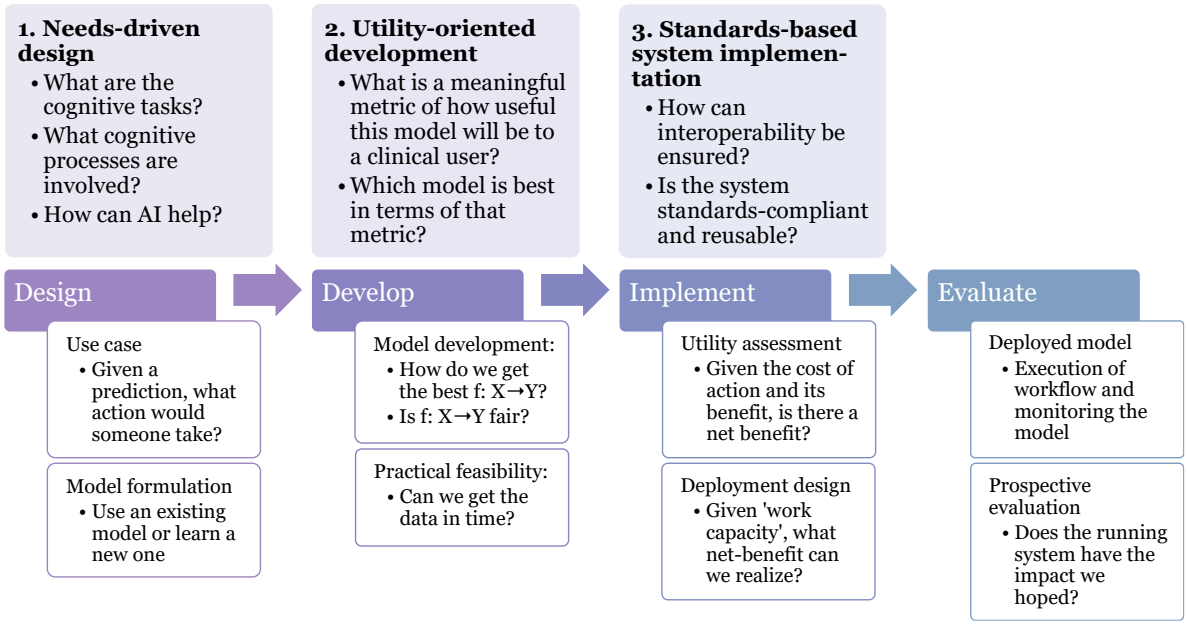

Figure 9.1. An extended framework for making predictive models useful in practice, with three additions (1-3)

9.1 Needs-driven design

First, I propose the inclusion of a systematic assessment of the clinical use case and corresponding user needs, including for clinical cognition and decision making, in collaboration with stakeholders. Specifically, cognitive engineering [29,71,269] approaches should be employed to understand the needs and design solutions for augmented AI components. The

existing framework emphasizes the use of human-centered design methods to understand user needs in the context of the clinical environment, and to develop and evaluate systems in the context of the clinical problem to be solved. However, both Li [21] and Jung [15] presume a use case focused on machine learning models that output risk predictions, exploring issues around making such risk scores clinically useful. Although risk scoring is an important area for AI, there are opportunities for AI to improve clinical care beyond such predictive modeling [270]. As Shneiderman argues, augmented intelligence, rather than automation, is the approach more suitable for developing useful, commercially viable applications [55,63,79]. This is because useful applications take a wide range of forms, many of which will not aim to automate classification tasks conventionally completed by human experts [55]. Additionally, this approach was expressly recommended in a special publication by the National Academic of Medicine on AI applications for health [7].

Cognitive engineering is the field concerned with augmenting human performance through computer assistance by supporting human cognitive processes [29]. Cognitive engineering approaches can be employed to characterize the cognitive tasks practitioners face as part of their work (such as diagnostic reasoning and care planning), as well as the cognitive processes involved in completing them(such as problem-solving, judgment, decision making, attention, perception, and memory). This approach thus illuminates possible avenues for AI to support the performance of medical practitioners. For example, the memory burden imposed by cross-referencing multiple sources of information and simultaneously filtering out irrelevant details may prove to be a significant cognitive bottleneck in formulating a clinical decision [271]. As a possible solution, the system might collect, summarize, filter, and present relevant information, such that it directly supports the processes of clinical comprehension and problem solving employed by clinical experts. For example, systems might develop and present intermediate constructs, or knowledge structures formed by experts as part of clinical cognition, such as clusters of symptoms that may point to a pattern indicative of a certain diagnosis [272].

This way, the system empowers human experts to focus their cognitive capacities on analyzing data and making clinical judgments and decisions by lowering the cognitive load imposed by the required steps. Adler-Milstein and colleagues term this objective "wayfinding", contrasting it with directly providing conclusions (e.g. diagnoses), a strategy unlikely to be trusted because it fails to acknowledge the complexity of the clinical task [81].

9.1.1 Application of needs-driven design in the current work

In this work, I employed needs assessment interviews to elucidate challenges and opportunities for informatics support within the Caring Contacts suicide prevention intervention (Chapter 3). I focused my investigation on two areas: first, contextual barriers to adoption, i.e. the constraints on the work system that provide the parameters within which any informatics tool operates; and second, challenges specific to individual workflow tasks. The second area concerns the individual tasks that intervention staff perform as part of their suicide prevention role and their mental models for completing them. I asked what problems needed to be solved and how they approached the solution. Further, I asked participants to share the most difficult parts of each task and describe in what ways they were demanding. I also asked them to share thoughts on, or experiences with, scenarios where those difficulties were alleviated. The challenges participants shared fell into five themes (Figure 9.2):

1. *Routine, time-consuming tasks* pose challenges on account of being labor-intensive, taking away time and focus from patient care.
2. *Managing risk* is a shared responsibility between stakeholders. Organizations should not promise services they do not have the capacity to provide. Vigilance is required regarding outgoing message content, monitoring and assessing risk in patient responses, and composing follow-up communications. Accurately assessing patient risk is paramount but can be complex. In the event of a patient response, clinicians consider not only overt message content, but also medical history, communication

style and personal attitudes, personal risk and protective factors, historical engagement with the healthcare system or prevention program, and more. Time pressure exacerbates the difficulty of this task, as swift and effective action is required to support a patient in an urgent crisis.

3. *Authoring follow-up messages* is sometimes trivial, but other times, it requires careful consideration. What follow-up message content is most useful will depend on individual patient factors. Staff may remind patients of the coping mechanisms they previously acknowledged as being helpful to them. If a safety plan is on file, staff may reiterate parts of that plan. Also, patients might face unique challenges that affect how staff can support them; for example, recommending a check-in with a behavioral health provider for a patient without access to care may be useless at best and harmful at worst.
4. *Accessing and integrating information across multiple data sources and information systems* incurs a workload burden and is cognitively demanding, especially when systems contain conflicting information.
5. *Team communication* is sometimes challenging, particularly when stakeholders work in different physical environments and information systems.

From these challenge-related themes, I developed design considerations for tools to support intervention staff in their day-to-day work.

First, I determined that fully or partially *automating* certain tasks should be a central focus for such tools. For example, in some programs, staff manually track message schedules each day, typing out or handwriting messages scheduled for that day. An information system could automatically schedule messages based on a template and send out messages according to that schedule, freeing up staff for other tasks. A system might also automatically assess the risk level expressed in incoming messages, alert staff of potentially urgent messages, and pre-sort

follow-up tasks in order of predicted priority, reducing the cognitive burden of assessing risk from scratch.

The second design focus should be on *data and workflow integration*. There should be a way for care team members to communicate about the patient in a centralized location or shared interface, rather than requiring users to log in to multiple systems. For example, intervention staff might access the system functions relevant to Caring Contacts from within the same electronic health record (EHR) chart that ongoing behavioral health providers use. Further, duplicated data entry should be relieved by enabling data exchange between systems, reducing the need to cross-reference multiple systems and reconcile conflicting information.

Third, tools should provide *cognitive support* for authoring follow-up messages. Aggregating patient data, filtering them for relevance, and presenting them in a digestible information display can help direct the attention of clinical staff, lightening cognitive burdens. A tool might also tag, extract, and highlight clinically relevant content from the medical history or the history of exchanged messages, such as risk and protective factors, which authors may want to incorporate. Offering relevant resources to include, e.g. crisis services located near the patient's geographic location, or suggesting revisions to in-progress messages to enhance them, for example to enhance empathy as demonstrated by Sharma et al. [140], also present avenues for cognitive support to help message authors increase message relevance and effectiveness.

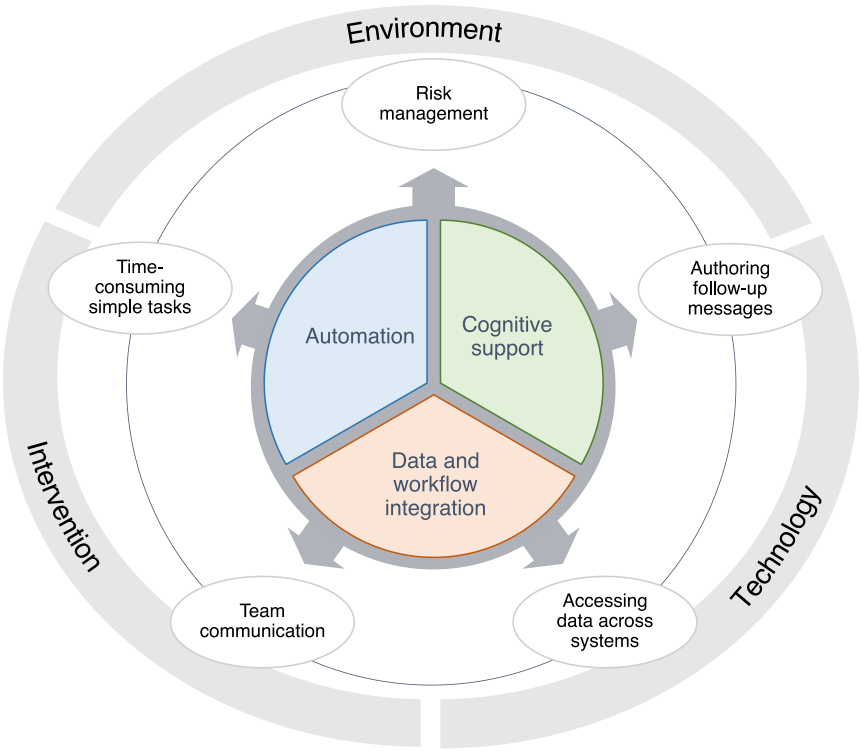

Figure 9.2 Summary of findings mapped to design considerations for informatics-supported suicide prevention.

Work system constraints are shown on the outer circle. Workflow challenges are shown in bubbles. Design considerations for addressing challenges are shown in the inner circle.

9.1.2 Conclusion regarding recommendations for needs-driven design

Much current machine learning work relevant to suicide prevention investigates suicide risk prediction, for example, based on EHR or social media data [92,150]. While I determined that there is a place for such suicide risk scoring models to support Caring Contacts, I also found a host of other opportunities for informatics support, using both AI-based and non AI-based approaches. The needs within the workflow are much broader than what appears to have been anticipated by machine learning researchers in the informatics and computer science communities. The best informatics tools for Caring Contacts will combine holistic workflow support functionality with various task-oriented, needs-driven AI components. Much current machine learning work relevant to suicide prevention investigates suicide risk prediction, for example, based on EHR or social media data [92,150]. While I determined that there is a place for such suicide risk scoring models to support Caring Contacts, I also found a host of other

opportunities for informatics support, using both AI-based and non AI-based approaches; AI-based opportunities included but went beyond prediction. The needs within the workflow are much broader than what appears to have been anticipated by machine learning researchers in the informatics and computer science communities. The best informatics tools for Caring Contacts will combine holistic workflow support functionality with various task-oriented, needs-driven AI components.

In agreement with contemporary proposals [55,56,80,81,273], these findings suggest that human-centered needs assessments that include elements focusing on clinical cognition have the potential to not only capture needs comprehensively but also to lend critical insights into what specific AI components may be helpful. Therefore, a generalized recommendation for a cognitive needs assessment is warranted.

9.2 Utility-oriented development

Throughout the design (1), development (2), implementation (3), and evaluation (4) stages for AI in healthcare, effectiveness should be evaluated. Evaluations will happen at various points along the spectrum from in-silico to in-vivo, in terms of metrics increasingly reflective of clinical impact. Jung and colleagues describe utility evaluations at the implementation stage (stage 3), after model development and validation (stage 2) are complete, to guide deployment strategies. However, if it becomes apparent that modifications to model conceptualization, design, or development are necessary, downstream revisions will be required. Shifting estimations of in-vivo effectiveness to earlier stages in this process will require fewer downstream revisions, preventing the waste of resources later on. This is in line with the common wisdom in the software engineering discipline: bugs found during the requirements gathering process, design, development, deployment, and maintenance phases are each exponentially more expensive to fix than those found earlier in this sequence.

As Jung et al. [15] demonstrated, workflow factors will critically affect utility and should therefore be estimated and incorporated when assessing the practical benefits of deploying the model for clinical use. Standardized system-centric metrics of predictive model performance, e.g. AUROC or the F1-score, are critical to compare performance across models but neglect to consider the impact of workflow parameters. In a clinical deployment, models with low performance scores might have a significant practical impact; conversely, models with high AUROC and F1 scores may have a negligible impact. Impact will depend on the aspects of the workflow that the model aims to modulate and the extent to which such modulation is possible. For example, rather than predictive accuracy, the true test of utility might be how many minutes can be saved in a particular workflow. If time requirements are determined mainly by factors outside the model's purview, even a high-performing model will not have a high impact. On the other hand, if the model critically impacts time requirements for every correct prediction, benefits might accumulate to substantial time savings even in the face of low sensitivity.

Therefore, it is vital to consider workflow constraints as early as possible, to test whether the model is performant enough to impact the chosen workflow at the designated workflow step. If model developers articulate performance in workflow-relevant terms that are intuitively interpretable, stakeholders can make well-informed decisions regarding the project's next steps. If impact cannot be expected confidently, informatics teams may choose to reassess machine learning approaches, workflow choices, and integration points within those workflows. Hence, I propose that clinical utility estimations are incorporated even at model development time (stage 2) before moving on to deployment planning (stage 3).

9.2.1 Application of utility-oriented development in the current work

In Chapter 3, I assessed the context of use and user needs within Caring Contacts and found that manual message triage is prohibitively labor-intensive. Accurate, timely triage is essential to avoid patient harm, but low resource availability limits staffing. Leveraging human

experts for triage is therefore not feasible for large-scale Caring Contacts. Automated triage of incoming patient messages for follow-up would reduce staffing requirements, facilitating cost-effect scaling of the intervention. In the work described in Chapter 6, I therefore developed and evaluated a machine learning model for classifying message urgency. The goal of the model is to alleviate potential delays in message response time in cases where a timely response is critical, but human triage is unavailable. The model's performance should be measured in the context of this goal, i.e. assigning relative priorities for clinician follow-up such that urgent messages can be addressed first. The model's potential for clinical utility should be measured in intuitively interpretable, workflow-relevant terms, i.e. the average response time for urgent messages. I compared several conditions:

- Absence of triage, i.e. messages are addressed in random order. This represents a setting where human triage is not available, for example, due to program scale.
- Automated triage. This represents a setting where an (imperfect) AI component is used to prioritize messages.
- Human triage. This represents what might be achieved in the ideal, but poorly scalable setting where human experts complete all triage.

The first and third conditions provide a lower and upper bound on the average response time, respectively. In order to have any utility at all, a model must outperform random triage. The performance target is that of human triage. However, that automated triage will not match human performance levels should be expected; thus, the reduction in staffing requirements afforded by automated triage comes at the expense of accuracy. Candidate models should be assessed and compared to determine the model that minimizes this need for compromise.

To accomplish this, I first estimated the workflow factors constraining response time. Using data collected as part of a clinical trial of two-way, text message-based Caring Contacts, and given the work capacity of research staff in this clinical trial, I estimated the potential response time that could be achieved for urgent messages, and the response time expected for

messages not considered urgent. These anticipated response times are invariants estimated from real-world data. In practice, they might also be provided by subject matter experts. I then devised a formula for determining the *average response time in urgent messages (ATRIUM)*, in minutes, over a range of work capacities, given triage performance. The work capacity is a parameter that might guide deployment strategies akin to Jung; therefore, it is important to estimate response time over a range of work capacities.

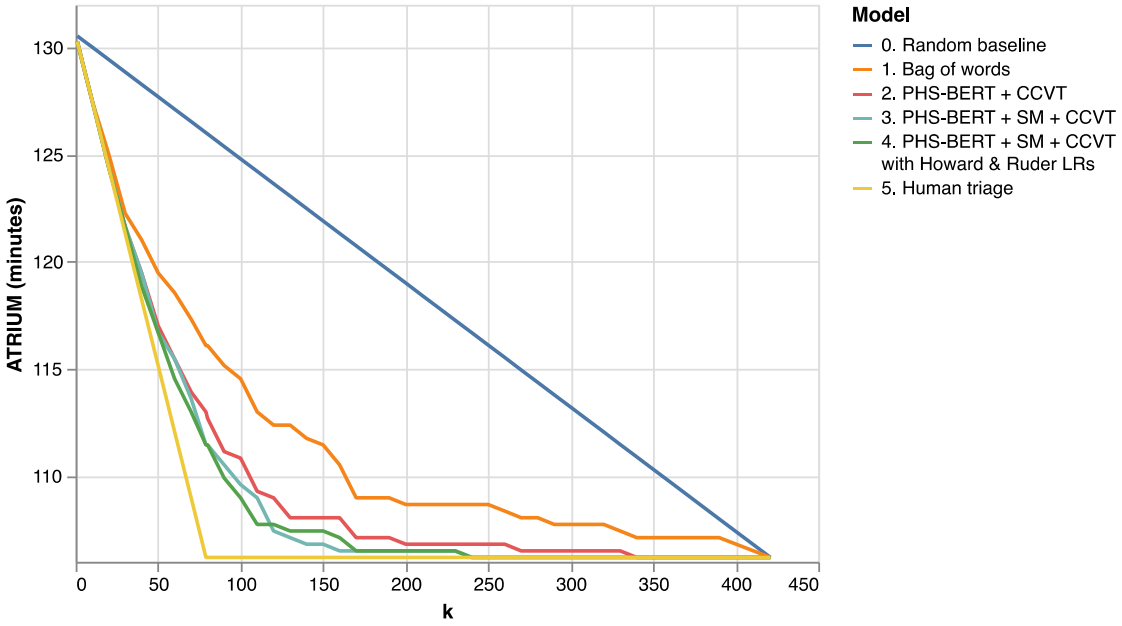

Figure 9.3 Average time to response in urgent messages (ATRIUM) for different triage approaches, given work capacity *k*, i.e. the number of messages that can be addressed with urgency

If work capacity is assumed to be equivalent to that in the clinical trial (arguably a best-case scenario on account of the availability of independent funding to support it), I determined that the following response times should be expected (Figure 9.3). If Caring Contacts were deployed without triage – either human or machine – urgent messages would be expected to receive a response after 126.4 minutes, on average. In contrast, human triage resulted in an estimated average response time of 106.2 minutes for urgent messages. I developed and evaluated several machine learning approaches, resulting in average expected response times for urgent messages ranging from 116.1 minutes to 111.6 minutes. In other words, human triage saved about 20 minutes per urgent message compared to responding randomly, and machine

triage saved between 10 and 15 minutes per message, depending on the machine learning techniques employed. The best model achieved an average response time within 5 minutes of that estimated with human triage.

These results demonstrate that machine learning can support the scaling of the intervention with minimal compromise to response time. I articulate model performance in workflow-relevant terms: Clinical stakeholders can use figures such as “10 minutes” and “20 minutes” to clearly visualize if the model meaningfully impacts care provision for a person facing a suicidal crisis and seeking urgent support. Additionally, clinical stakeholders can use this metric to design a deployment environment that maximizes clinical utility based on model performance: Clinical stakeholders can choose a staffing level that results in a value for the work capacity k that results in an acceptable tradeoff between staffing requirements and response times. This is akin to Jung, who discussed using utility calculations across different model configurations to inform deployment designs that optimize utility; namely, they compared the effects of increasing inpatient vs. outpatient capacity given model performance.

9.2.2 Conclusion regarding recommendations for utility-oriented development

My findings demonstrate the feasibility of estimating clinical utility at model development time. The true clinical impact of the model depends on the deployment conditions and is, therefore, unknowable at development time. However, utility estimations performed at this stage can inform implementation decisions better than AUROC or F1 scores, which may be not only uninterpretable for clinical stakeholders, but also meaningless without consideration of workflow factors. This work has provided one example of a development-time utility metric, but priorities will vary across problems, and factors such as net financial benefits may be more pertinent than time constraints in other applications. Other examples include time utility (via hierarchical TBG) as demonstrated by Shing et al. [142], and cost utility (in US Dollars) as

demonstrated by Jung et al. [15] and Bayati et al. [22]. Thus, development-time utility evaluation is valuable, and its inclusion in a generalized framework is warranted.

9.3 Standards-based system implementation

Data and messaging standards are now widely recognized as keystones for achieving the vision of the learning health system, which recently culminated in the 21st Century Cures Act mandating that healthcare organizations implement and maintain standards-based application programming interfaces (APIs) [145,146]. Currently, many digital health and healthcare AI "delivery systems" do not make use of these standards: standalone, purpose-built solutions such as LAMP [259] are common, as are systems that make use of EHR vendors-specific capabilities to run predictive models or integrated predictions back into the workflow, e.g. the VSAIL model for predicting suicide risk from structured EHR data [118]. While vendor support at least enables some knowledge sharing amongst organizations using the same vendor, organizations may still have to design and develop their own models, workflows, and integration strategies for CDS. Solutions that are interoperable with existing systems and that can be reused across implementation sites with minimal duplication of work will accelerate the widespread, equitable adoption of AI in healthcare. Sharing work in this way is integral to allowing all organizations to benefit from healthcare AI, even those who do not have the expertise or funds to design and develop their own. The implementation of sustainable, interoperable AI-based information systems should therefore be prioritized.

Fortunately, standards-based software development paradigms, standardized vocabularies and ontologies, CDS-specific technologies, and best practice operating procedures are available and finding broader adoption [159]. I propose the inclusion of considerations for portability, interoperability, and standardization. Which approaches and technologies are appropriate will depend on the use case but may include the following:

- standardized vocabularies, e.g. International Classification of Diseases (ICD) [274], Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [275], and Logical Observation Identifiers Names and Codes (LOINC) [276]
- data model and information exchange standards, such as Health Level 7 (HL7) Version 3 [277] and HL7 Fast Healthcare Interoperability Resources (FHIR) [26]
- standardized clinical knowledge expression approaches, such as the Arden Syntax [278] and Clinical Quality Language (CQL) [279]
- workflow integration protocols, including Substitutable Medical Applications and Reusable Technologies (SMART) -on-FHIR [25] and CDS Hooks [280]
- reproducibility considerations such as code, data, and model sharing

9.3.1 Application of standards-based system implementation in the current work

In Chapter 3, I assessed barriers to adoption and opportunities for technology support within Caring Contacts and found that data and workflow integration represented a significant opportunity to address challenges. Patient data were reported to be spread across multiple systems, requiring duplicative data entry, cognitive effort to consolidate information from different sources, and multiple logins. Overall, participants reported variations in workflow, but challenges and needs had common threads across all programs represented by participants, suggesting the feasibility of a shared solution. In Chapter 8, I therefore conceptualized, implemented, and shared a standards-based information system called Informatics-Supported Administration for Caring Contacts (ISACC).

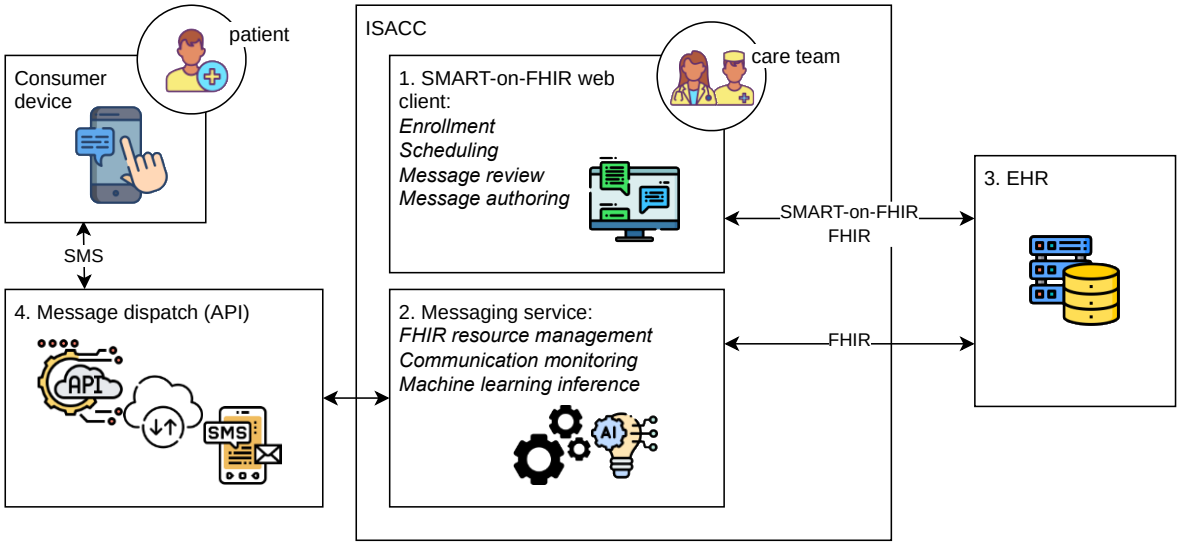

Figure 9.4. ISACC application architecture

To accomplish this, I first developed an application requirements specification for ISACC based on the design considerations set in Chapter 3. I then assessed the corresponding data model requirements and developed a FHIR representation scheme to meet these requirements. Next, I devised a FHIR-based system architecture (Figure 9.4) that supports the application requirements, including real-time classification of patient-generated text messages for urgency using the machine learning model developed in Chapter 6. Finally, I developed and shared a functional open-source software application embodying these ideas.

FHIR enables interoperability via a shared data model with a well-defined specification that health information technology vendors are now required to implement. In ISACC, data and workflow integration is achieved via SMART-on-FHIR, which allows database and user interface sharing between the EHR and ISACC, eliminating the disadvantages of multiple databases and credentialing systems. This standardization also enables portability between different health systems: ISACC is reusable at other implementation sites without requiring customizations to data structures and exchange protocols. ISACC uses controlled vocabularies and value sets, supporting semantic interoperability, and uses FHIR resources to store and track machine

learning inferences. Finally, in the spirit of reproducibility, transparency, and reusability, ISACC is an open-source application for use or adaptation by informatics researchers and developers.

ISACC's approach to the integration of AI components facilitates AI adoption as well as clinical utility. A core challenge with patient-generated data (PGD) is their raw and variable nature. PGD are often ample, requiring extensive review and synthesis to determine any clinically actionable signal within, and thus may not be useful to clinicians operating under time constraints. This is especially true for natural language data, which would require clinicians to read and assess each document carefully to determine clinical relevance. Thus, to realize the potential of PGD, they should be processed before they can be effectively integrated at the point of care. ISACC accommodates AI components that infer clinically meaningful insights from incoming PGD, translates them into FHIR resources of a well-defined, standardized form, and deposits them in a shared database. Secondary use of insights like suicide risk scores, e.g. for reporting or research, is also enabled by this database sharing. Although organizations may tailor models to ensure high performance for their unique patient population, the infrastructure for using these models and the integration of outputs into workflows in clinically useful ways is shared. It does not have to be redesigned from scratch.

I thus demonstrated the feasibility of using FHIR and SMART-on-FHIR as the basis for a suicide prevention information system that leverages PGD and machine learning to support clinical workflows, enabling out-of-the-box interoperability and portability.

9.3.2 Conclusion regarding standards-based system implementation

Workflow integration issues, which are challenging because of expertise, funding, variability in data, and more, limit the successful and effective use of AI in clinical practice. My findings demonstrate the feasibility of shareable data and workflow integration, reducing the need to solve this complex problem repeatedly. This suggests that standards-based system design and development can facilitate the development and adoption of clinically useful AI.

Using standards to design for interoperability, portability, and reusability is a core requirement for new health information technology, and should be for AI-based clinical decision support systems, as well. Therefore, this general recommendation is warranted.

Chapter 10. Discussion & Conclusion

Artificial intelligence (AI) holds significant promise to improve health and healthcare. The reasons why this promise has not come to fruition are complex, with challenges at every step along the continuum from design to development, implementation, and evaluation. In this work, I investigated solutions to several of these challenges and presented the larger implications of my findings. However, overcoming these challenges will require significant efforts over the coming years and decades – this work represents a small piece of the puzzle.

In this dissertation, I investigated how needs-driven design, utility-oriented model development and evaluation methods, and standards-based software can be leveraged collectively to address the unique challenges faced by healthcare AI, and achieve clinically impactful AI implementations.

In my first aim (Chapter 3), I used human-centered design methods to assess technological support needs for Caring Contacts, an evidence-based suicide prevention intervention. Using stakeholder interviews and qualitative analysis, I characterized the challenges hampering the broad adoption of Caring Contacts and elucidated opportunities for technology support. I found that designers of digital tools should focus on automating routine, labor-intensive tasks, integrating data and workflows for improved workflow efficiency, and providing cognitive support for clinical follow-up tasks. In my second aim (Chapter 6), I developed a predictive model of suicide risk from patient-generated natural language and a metric for evaluating its clinical utility. I found that leveraging social media data for neural transfer learning significantly enhances predictive accuracy for risk-based prioritization of patient messages. I formulated a utility-oriented evaluation metric articulating clinical utility in workflow-relevant terms, the *average time to response in urgent messages (ATRIUM)*. I demonstrated that this model has the potential to impact clinical practice positively. In my third aim (Chapter 8), I developed blueprints for a standards-based, reusable, interoperable,

workflow-integrated information system for workflow and cognitive support of Caring Contacts. I then implemented and shared an open-source software application embodying these blueprints, demonstrating the feasibility of an AI-based system using patient-generated data (PGD) for clinical decision support (CDS).

10.1 Contributions

A framework for the needs-driven, utility-oriented, standards-based operationalization of AI for healthcare. The overarching contribution of this work is a generalizable framework for the operationalization of AI to support workflows and clinical decision making in healthcare. Compared to prior frameworks for clinical AI implementations, the specific additions are:

- Needs-driven design: Use cognitive needs assessments to discover opportunities for AI and inform the design of AI components.
- Utility-oriented development: Define use case-specific utility metrics and use them as part of AI development.
- Standards-based implementation: When implementing AI-based tools, utilize appropriate standards for portability and interoperability.

These recommendations address specific key challenges that have hampered the adoption of AI, contributing toward resolving the historical implementation gap of healthcare AI.

Application of this framework to conceive, implement and evaluate AI support for suicide prevention. Several contributions result from applying this framework toward AI-supported information technology for the Caring Contacts suicide prevention intervention.

In aim 1, I contribute an improved understanding of how suicide prevention interventions are conducted in practice, and the clinical cognition involved. While the underutilization of Caring Contacts was known, my work illuminated the specific reasons for this. I identified previously unrecognized opportunities for technological support, contributing to the path forward with design considerations for AI-enabled digital tools for this intervention. While human-centered design methods are broadly used to assess needs and design digital health tools, I specifically demonstrated the vital role cognitive needs assessments play in AI design and development.

In aim 2, I develop and evaluate natural language processing (NLP) approaches to suicide risk prediction in a clinical dataset using neural transfer learning. Publicly available social media data has been used for suicide prediction, but the applicability of such work to prediction tasks in clinical settings was previously unknown. I established the utility of publicly available social media benchmark data for downstream tasks involving patient-generated data from a clinical setting, and showed that contemporary techniques that facilitate the transfer of linguistic information further enhance this utility. Thus, aim 2 contributes towards addressing the limited availability of clinical data for machine learning. Aim 2 also contributes a novel metric of clinical utility. While time and cost utility metrics have been developed and used for other purposes, use case-specific, workflow-relevant metrics are needed to establish the utility of using AI to support. I contribute a metric that provides stakeholders with evidence of practical impact that is more relevant and interpretable than conventional measures of predictive performance.

In aim 3, I develop a standards-based information system incorporating patient-generated data (PGD) and AI for workflow and clinical decision support at the point of care. While Fast Healthcare Interoperability Resources (FHIR) is now widely used, no guidance for leveraging FHIR for text message-based suicide prevention is available. I devised a FHIR-native data representation model for AI-supported Caring Contacts, providing a starting point for

applications with similar data representation needs and contributing towards semantic interoperability of such applications. Further, the SMART-on-FHIR paradigm holds promise for applications with PGD-based AI, but the specifics of how to apply this technology to this case are not well understood. I developed an application architecture that continually takes in patient-generated data, infers clinically relevant insights, and uses SMART-on-FHIR for workflow integration, and demonstrated its feasibility through an implemented tool, contributing toward a better understanding of how to use SMART-on-FHIR in conjunction with PGD. Aim 3 further contributes a freely available software artifact illustrating these blueprints.

Finally, this work collectively makes an important contribution to suicide prevention. Caring Contacts is an example of an intervention that stands to benefit from AI, but the many challenges described throughout this work complicate the realization of this benefit. In investigating these challenges in the context of Caring Contacts, I produced solutions directly applicable to this intervention. The Informatics-Supported Administration of Caring Contacts (ISACC) application is the result of these efforts and represents an important contribution towards facilitating broader adoption of this effective but under-utilized suicide prevention intervention.

10.2 Generalizability

In this work, I attempted to develop a generalizable framework for AI in healthcare, but health informatics is a heterogeneous field encompassing multifaceted methods and applications. There are, therefore, limitations to the applicability of this work.

Here, I have focused on patient-generated natural language data. Other types of data, both internal and external to the Electronic Health Record (EHR), can serve as the basis for AI, e.g. structured PGD such as sensor data from wearables, patient-reported outcomes data (PROs) entered via (EHR-internal) questionnaires, public health data such as from opioid prescription databases, clinical notes, and structured EHR data such as medications, diagnoses, and vital

signs. My recommendations are not directly dependent on the type of data used, but the relative importance of the challenges they address may vary. Additionally, different data types will precipitate additional challenges and may require unique approaches. Specifically, EHR data, a valuable source of evidence commonly used as the basis for AI-powered CDS, differ in many respects from PGD. Though the recommendations in the novel framework are applicable, the challenges encountered for such CDS systems will differ from those in PGD-based systems.

Additionally, I focused on individual users' cognitive needs in this work. Other nuanced approaches to capturing needs may be appropriate depending on the setting. For example, in collaborative or team-based settings, which are common in clinical work, adopting a higher-level sociotechnical perspective may reveal additional barriers and opportunities for computer-supported co-operative work.

This work is concerned with suicide prevention. There are unique considerations with this problem space. For example, suicidal crises are comparatively rare, but exceptionally high-impact events. In the U.S. healthcare system, prevention interventions face cost-related challenges not shared by interventions with costs that are immediately reimbursable by payors: Cost utility largely depends on intervention costs and insurance reimbursement considerations. Machine learning techniques and evaluation approaches may vary for different problems; for example, publicly available data to counteract the limitations of clinical data may not be available, or workflow parameters needed for development-time utility estimations may not be available. Considerations for workflow integration differ between health problems and deployment sites. Although the recommendations in the framework are intended to guide informaticians in finding scenario-specific solutions, I validated the framework in the context of suicide prevention; consequently, other application areas may reveal the need for further recommendations.

The framework's evaluation step (step 4) was not included in this work for reasons of scope. Consequently, recommendations for this phase, beyond previously published guidance,

cannot be made or evaluated on the basis of this work. However, as the ultimate test of system success, evaluation is an indispensable part of developing effective AI-based CDS systems. Future work will evaluate the impact of ISACC in a range of clinical settings.

Intervention design and, as a result, workflow integration considerations may vary significantly between implementation sites. I assessed needs comprehensively by interviewing participants with different roles, representing a range of organizations, and serving diverse patient populations. Understanding that implementing organizations' needs and preferences will differ, I designed and developed ISACC with configurability in mind. Yet, customizations of varying degrees will be required before deployment at different sites can be completed. Further development may be necessary, e.g. if organizations wish to use ISACC for email or postal mail-based interventions.

Finally, the broader adoption of Caring Contacts will depend on factors outside the control of any informatics tool. While tools should be developed to account for contextual constraints, there are reasons why even the best tools will not be adopted. For example, even with reduced resource requirements, some organizations will not have the capacity for a suicide prevention program. Other health issues may be prioritized. Technology availability and literacy issues may preclude technological interventions altogether, for example, in low-income countries or within certain patient populations.

10.3 Future work

Future work may apply the framework to different settings, revealing how the framework can be expanded or refined further to account for the limitations mentioned in the preceding section. Additionally, future work that includes the evaluation phase may reveal opportunities to expand the framework and make it more useful for this phase.

Aim 1 revealed a range of opportunities for AI support, including mining message histories for relevant, clinically actionable content. While Chapter 4 and Chapter 5 presented

preliminary work to this end, further work is needed to fully develop such AI components and realize these opportunities for AI support for Caring Contacts. Similarly, aim 1 revealed opportunities for information displays incorporating such extracted information as well as EHR data. Such information displays represent an important opportunity for cognitive support. Future work may include iterative design and evaluation of such information displays using human-centered design methods.

An extensive body of work explores the use of structured EHR data for suicide prediction, with promising results. Future work could investigate multimodal suicide risk prediction based on both patient-generated natural language and structured EHR data to improve prediction performance further.

Concerning Caring Contacts, future work entails developing ISACC further and ensuring its applicability to individual deployment settings. Extensive usability and integration testing will be needed to ready the application for pilot testing and eventual deployment.

10.4 Conclusion

In this dissertation, I identified several key challenges to the adoption of AI in healthcare, as well as gaps in existing guidance for addressing these challenges, and developed a framework that fills these gaps. I evaluated the framework in the context of suicide prevention with Caring Contacts and found that my additions resulted in an improved understanding of the problem and a clearer formulation of solutions, an evaluation approach that more closely reflects real-world value, and an information system that lowers important barriers associated with dissemination and deployment. Though a small piece in the puzzle, this work contributes towards bridging the historical implementation gap by furthering methods for design, development, and delivery of AI-supported interventions, and by guiding future attempts to realize the potential of AI in clinical settings.

References

1 OECD. Health spending (indicator). 2022. doi:doi: 10.1787/8643de7e-en

2 OECD. Life expectancy at birth (indicator). 2022. doi:10.1787/27e0fc9d-en

3 Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ* 2016;**353**. doi:10.1136/bmj.i2139

4 Institute of Medicine. *Who Will Keep the Public Healthy?* Washington, DC: : National Academies Press 2003. doi:10.17226/10542

5 Berwick DM, Nolan TW, Whittington J. The triple aim: Care, health, and cost. *Health Aff* 2008;**27**:759–69. doi:10.1377/hlthaff.27.3.759

6 Bodenheimer T, Sinsky C. From triple to Quadruple Aim: Care of the patient requires care of the provider. *Ann Fam Med* 2014;**12**:573–6. doi:10.1370/afm.1713

7 National Academy of Medicine. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. 2019.

8 Friedman CP, Wong AK, Blumenthal D. Policy: Achieving a nationwide learning health system. *Sci Transl Med* 2010;**2**:1–4. doi:10.1126/scitranslmed.3001456

9 Sheikh A, Sood HS, Bates DW. Leveraging health information technology to achieve the 'triple aim' of healthcare reform. *J Am Med Informatics Assoc* 2015;**22**:849–56. doi:10.1093/jamia/ocv022

10 Friedman CP, Rubin J, Brown J, *et al*. Toward a science of learning systems: A research agenda for the high-functioning Learning Health System. *J Am Med Informatics Assoc* 2015;**22**:43–50. doi:10.1136/amiajnl-2014-002977

11 Shortliffe EH. The adolescence of AI in Medicine: Will the field come of age in the '9Os?*. 1993.

12 Patel VL, Shortliffe EH, Stefanelli M, *et al*. The coming of age of artificial intelligence in medicine. *Artif Intell Med* 2009;**46**:5–17. doi:10.1016/j.artmed.2008.07.017

13 Shortliffe EH. Artificial Intelligence in Medicine: Weighing the Accomplishments, Hype, and Promise. *Yearb Med Inform* 2019;**28**:257–62. doi:10.1055/s-0039-1677891

14 Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innov. 2020;**6**:45–7. doi:10.1136/bmjinnov-2019-000359

15 Jung K, Kashyap S, Avati A, *et al*. A framework for making predictive models useful in practice. *J Am Med Informatics Assoc* 2020;**00**:1–10. doi:10.1093/jamia/ocaa318

16 Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA* 2019;**321**:2281. doi:10.1001/jama.2019.4914

17 Shah NH, Milstein A, Bagley, PhD SC. Making Machine Learning Models Clinically Useful. *JAMA* 2019;**322**:1351. doi:10.1001/jama.2019.10306

18 Wainberg M, Merico D, Delong A, *et al*. Deep learning in biomedicine. *Nat Biotechnol* 2018;**36**:829–38. doi:10.1038/nbt.4233

19 Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* Published Online First: 2018. doi:10.1098/rsif.2017.0387

20 Watson J, Hutyra CA, Clancy SM, *et al*. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020;**3**:167–72. doi:10.1093/jamiaopen/ooz046

21 Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. npj Digit. Med. 2020;**3**. doi:10.1038/s41746-020-00318-y

22 Bayati M, Braverman M, Gillam M, *et al*. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One* 2014;**9**:1–9. doi:10.1371/journal.pone.0109264

23 Burkhardt H, Brandt P, Lee J, *et al*. StayHome: A FHIR-Native Mobile COVID-19 Symptom Tracker and Public Health Reporting Tool. *Online J Public Health Inform* 2021;**13**. doi:10.5210/ojphi.v13i1.11462
24 Jenders RA, Del Fiol G, Haug P, *et al*. Health Information Technology Standards for Implementing and Using Clinical Decision Support : Latest Developments and What You Need to Know. In: *Annual Symposium of the Journal of the American Medical Informatics Association*. 2019.
25 Mandel JC, Kreda DA, Mandl KD, *et al*. SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *J Am Med Informatics Assoc* 2016;**23**:899–908. doi:10.1093/jamia/ocv189
26 Bender D, Sartipi K. HL7 FHIR: An agile and RESTful approach to healthcare information exchange. In: *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013. 326–31. doi:10.1109/CBMS.2013.6627810
27 Chute CG, Koo D. Public health, data standards, and vocabulary: Crucial infrastructure for reliable public health surveillance. *J Public Heal Manag Pract* 2002;**8**:11–7. doi:10.1097/00124784-200205000-00003
28 Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 2018;**320**:2199. doi:10.1001/jama.2018.17163
29 Roth EM, Patterson ES, Mumaw RJ. Cognitive engineering: Issues in user-centered system design. *Encycl Softw Eng 2nd Ed* 2001;:2076.
30 Stone DM, Simon TR, Fowler KA, *et al*. Trends in state suicide rates 1999-2016. *Morb Mortal Wkly Rep* 2018;**67**:617–24.
31 Substance Abuse and Mental Health Services Administration(SAMHSA). Key Substance Use and Mental Health Indicators in the United States: Results from the 2020 National Survey on Drug Use and Health. *HHS Publ No PEP19-5068, NSDUH Ser H-54* 2021;**170**:1–62.https://www.samhsa.gov/data/
32 Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2019 National Survey on Drug Use and Health. Rockville, MD: 2020. https://www.samhsa.gov/data/
33 Underlying Cause of Death 1999-2020 on CDC WONDER Online Database. http://wonder.cdc.gov/ucd-icd10.html
34 Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: A review of the evidence. *Am J Psychiatry* 2002;**159**:909–16. doi:10.1176/appi.ajp.159.6.909
35 Simon GE, Yarborough BJ, Rossom RC, *et al*. Self-reported suicidal ideation as a predictor of suicidal behavior among outpatients with diagnoses of psychotic disorders. *Psychiatr Serv* 2019;**70**:176–83. doi:10.1176/appi.ps.201800381
36 The Joint Commission. National Patient Safety Goal for suicide prevention. *R3 Rep* 2019;:1–6.
37 Stanley B, Brown GK, Brenner LA, *et al*. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. *JAMA Psychiatry* 2018;**75**:894–900. doi:10.1001/JAMAPSYCHIATRY.2018.1776/
38 Luxton DD, June JD, Comtois KA. Can postdischarge follow-up contacts prevent suicide and suicidal behavior? A Review of the Evidence. *Crisis* 2013;**34**:32–41. doi:10.1027/0227-5910/a000158
39 Motto JA, Heilbron DC, Juster RP, *et al*. Communication as a Suicide Prevention Program. In: *Depression and Suicide*. Elsevier 1983. 148–54. doi:10.1016/B978-0-08-027080-7.50031-8
40 Comtois KA, Kerbrat AH, DeCou CR, *et al*. Effect of Augmenting Standard Care for Military Personnel With Brief Caring Text Messages for Suicide Prevention. *JAMA*

Psychiatry 2019;**76**:474. doi:10.1001/jamapsychiatry.2018.4530
41 Reger MA, Luxton DD, Tucker RP, *et al*. Implementation methods for the caring contacts suicide prevention intervention. *Prof Psychol Res Pract* 2017;**48**:369–77. doi:10.1037/pro0000134
42 U.S Department of Veteran Affairs. VA/DoD Clinical Practice Guideline for the Assessment and Management of Patients at Risk of Suicide. 2019;:35.https://www.healthquality.va.gov/guidelines/MH/srb/
43 Landes SJ, Jegley SM, Kirchner JE, *et al*. Adapting Caring Contacts for Veterans in a Department of Veterans Affairs Emergency Department: Results From a Type 2 Hybrid Effectiveness-Implementation Pilot Study. *Front Psychiatry* 2021;**12**:1–11. doi:10.3389/fpsyt.2021.746805
44 Skopp NA, Smolenski DJ, Bush NE, *et al*. Caring contacts for suicide prevention: A systematic review and meta-analysis. *Psychol Serv* Published Online First: 14 April 2022. doi:10.1037/ser0000645
45 Landes SJ, Kirchner JAE, Areno JP, *et al*. Adapting and implementing Caring Contacts in a Department of Veterans Affairs emergency department: A pilot study protocol. *Pilot Feasibility Stud* 2019;**5**:1–11. doi:10.1186/s40814-019-0503-9
46 Motto JA, Bostrom AG. A randomized controlled trial of postcrisis suicide prevention. *Psychiatr Serv* 2001;**52**:828–33. doi:10.1176/APPI.PS.52.6.828/ASSET/IMAGES/LARGE/HD21T4.JPEG
47 Luxton DD, Smolenski DJ, Reger MA, *et al*. Caring E-mails for Military and Veteran Suicide Prevention: A Randomized Controlled Trial. *Suicide Life-Threatening Behav* 2020;**50**:300–14. doi:10.1111/sltb.12589
48 Reger MA, Gebhardt HM, Lee JM, *et al*. Veteran Preferences for the Caring Contacts Suicide Prevention Intervention. *Suicide Life-Threatening Behav* 2019;**49**:1439–51. doi:10.1111/sltb.12528
49 Nelson L, Comtois K. Caring Contacts: A Strength-based, Suicide Prevention Trial in 4 Native Communities (CARE). clinicaltrials.gov. 2016.https://clinicaltrials.gov/ct2/show/NCT02825771
50 Comparing two ways for healthcare providers to intervene to prevent suicide among adults and adolescents. 2019.https://www.pcori.org/research-results/2019/comparing-two-ways-provide-safety-planning-follow-support-adults-and-teens-risk-suicide
51 Demszky D, Movshovitz-Attias D, Ko J, *et al*. GoEmotions: A Dataset of Fine-Grained Emotions. Published Online First: 1 May 2020.http://arxiv.org/abs/2005.00547
52 Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digit Med* 2020;**3**. doi:10.1038/s41746-020-0233-7
53 Guntuku SC, Yaden DB, Kern ML, *et al*. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci* 2017;**18**:43–9. doi:10.1016/j.cobeha.2017.07.005
54 Gruber J, Oveis C, Keltner D, *et al*. A discrete emotions approach to positive emotion disturbance in depression. *Cogn Emot* 2011;**25**:40–52. doi:10.1080/02699931003615984
55 Shneiderman B. Design Lessons From AI's Two Grand Goals: Human Emulation and Useful Applications. *IEEE Trans Technol Soc* 2020;**1**:73–82. doi:10.1109/tts.2020.2992669
56 Heer J. Agency plus automation: Designing artificial intelligence into interactive systems. *Proc Natl Acad Sci USA* 2019;**116**:1844–50. doi:10.1073/pnas.1807184115
57 Prisco J. A short history of the elevator. CNN Style. 2019.https://www.cnn.com/style/article/short-history-of-the-elevator/index.html (accessed 12 Nov 2022).

58 Vigliarolo B. Amazon Alexa: Cheat sheet. *TechRepublic* 2020.https://www.techrepublic.com/article/amazon-alexa-the-smart-persons-guide/ (accessed 12 Nov 2022).
59 Hoy MB. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med Ref Serv Q* 2018;**37**:81–8. doi:10.1080/02763869.2018.1404391
60 Rogers EM. *Diffusion of Innovations*. 1962.
61 Bansal G, Nushi B, Kamar E, *et al*. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *33rd AAAI Conf Artif Intell AAAI 2019, 31st Innov Appl Artif Intell Conf IAAI 2019 9th AAAI Symp Educ Adv Artif Intell EAAI 2019* 2019;:2429–37. doi:10.1609/aaai.v33i01.33012429
62 Friedman CP. A 'Fundamental Theorem' of Biomedical Informatics. *J Am Med Informatics Assoc* 2009;**16**:169–70. doi:10.1197/jamia.M3092
63 Shneiderman B. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *Int J Hum Comput Interact* 2020;**36**:495–504. doi:10.1080/10447318.2020.1741118
64 Hartzler AL, Bartlett LE, Hobler MR, *et al*. Take on transplant: human-centered design of a patient education tool to facilitate informed discussions about lung transplant among people with cystic fibrosis. *J Am Med Informatics Assoc* 2022;**00**:1–12. doi:10.1093/jamia/ocac176
65 Faiola A, Srinivas P, Duke J. Supporting Clinical Cognition: A Human-Centered Approach to a Novel ICU Information Visualization Dashboard. *AMIA . Annu Symp proceedings AMIA Symp* 2015;**2015**:560–9.
66 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J Am Med Assoc* Published Online First: 2016. doi:10.1001/jama.2016.17216
67 Keep an Eye on Your Vision Health | CDC. https://www.cdc.gov/visionhealth/resources/features/keep-eye-on-vision-health.html (accessed 8 Oct 2022).
68 National Center for Health Statistics. Percentage of having a wellness visit in past 12 months for adults aged 18 and over, United States, 2019—2020. Natl. Heal. Interview Surv. https://wwwn.cdc.gov/NHISDataQueryTool/SHS_adult/index.html (accessed 8 Oct 2022).
69 Gobbi JD, Braga JPR, Lucena MM, *et al*. Efficacy of smartphone-based retinal photography by undergraduate students in screening and early diagnosing diabetic retinopathy. *Int J Retin Vitr* 2022;**8**:1–9. doi:10.1186/s40942-022-00388-y
70 Beede E, Baylor E, Hersch F, *et al*. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Conf Hum Factors Comput Syst - Proc* 2020;:1–12. doi:10.1145/3313831.3376718
71 Woods DD, Roth EM. Cognitive Engineering: Human Problem Solving With Tools. *Hum Factors* 1988;**30**:415–30. doi:10.1177/001872088803000404
72 Norman DA. Cognitive engineering. In: *User Centered System Design*. Lawrence Erlbaum Association 1986. doi:10.1201/b15703-3
73 Hettinger AZ, Roth EM, Bisantz AM. Cognitive engineering and health informatics: Applications and intersections. *J Biomed Inform* 2017;**67**:21–33. doi:10.1016/j.jbi.2017.01.010
74 Klein GA, Calderwood R, Macgregor D. Critical Decision Method for Eliciting Knowledge. *IEEE Trans Syst Man Cybern* 1989;**19**:462–72. doi:10.1109/21.31053
75 Militello LG, Hutton RJB. Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *http://dx.doi.org/101080/001401398186108* 2010;**41**:1618–41. doi:10.1080/001401398186108
76 Xiao Y. Artifacts and collaborative work in healthcare: methodological, theoretical, and technological implications of the tangible. *J Biomed Inform* 2005;**38**:26–33.

doi:10.1016/J.JBI.2004.11.004
77 Norman DA. Chapter 3: The Power of Representation. In: *Things That Make Us Smart*. 1993.
78 Bauer DT, Guerlain SA, Brown PJ. Evaluating the use of flowsheets in pediatric intensive care to inform design. *Proc Hum Factors Ergon Soc* 2006;:1057–8. doi:10.1177/154193120605001011
79 Shneiderman B. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans Interact Intell Syst* 2020;**10**. doi:10.1145/3419764
80 Xu W. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* 2019;**26**:42–6. doi:10.1145/3328485
81 Adler-Milstein J, Chen JH, Dhaliwal G. Next-Generation Artificial Intelligence for Diagnosis. *JAMA* 2021;**326**:2467. doi:10.1001/jama.2021.22396
82 Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: An integrative review. *J Am Med Informatics Assoc* 2018;**25**:759–71. doi:10.1093/jamia/ocy006
83 Adler-Milstein J, Nong P. Early experiences with patient generated health data: Health system and patient perspectives. *J Am Med Informatics Assoc* 2019;**26**:952–9. doi:10.1093/jamia/ocz045
84 Austin E, Lee JR, Amtmann D, *et al*. Use of patient-generated health data across healthcare settings: implications for health systems. *JAMIA Open* 2020;**3**:70–6. doi:10.1093/jamiaopen/ooz065
85 Tiase VL, Hull W, McFarland MM, *et al*. Patient-generated health data and electronic health record integration: a scoping review. *JAMIA Open* Published Online First: 5 December 2020. doi:10.1093/jamiaopen/ooaa052
86 Cohen DJ, Keller SR, Hayes GR, *et al*. Integrating Patient-Generated Health Data Into Clinical Care Settings or Clinical Decision-Making: Lessons Learned From Project HealthDesign. *JMIR Hum Factors* 2016;**3**:e26. doi:10.2196/humanfactors.5919
87 Woods SS, Evans NC, Frisbee KL. Integrating patient voices into health information for self-care and patient-clinician partnerships: Veterans Affairs design recommendations for patient-generated data applications. *J Am Med Informatics Assoc* 2016;**23**:491–5. doi:10.1093/jamia/ocv199
88 Babiarczyk B, Sternal D. Accuracy of self-reported and measured anthropometric data in the inpatient population. *Int J Nurs Pract* 2015;**21**:813–9. doi:10.1111/ijn.12314
89 Campbell RS, Pennebaker JW. The Secret Life of Pronouns. *Psychol Sci* 2003;**14**:60–5. doi:10.1111/1467-9280.01419
90 Pennebaker JW. *The secret life of pronouns: What our words say about us*. 1st U.S. e. New York: : Bloomsbury Press 2011.
91 Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annu Rev Psychol* 2003;**54**:547–77. doi:10.1146/annurev.psych.54.101601.145041
92 Resnik P, Foreman A, Kuchuk M, *et al*. Naturally occurring language as a source of evidence in suicide prevention. *Suicide Life-Threatening Behav* 2020;:1–9. doi:10.1111/sltb.12674
93 Coppersmith G, Leary R, Crutchley P, *et al*. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomed Inform Insights* 2018;**10**:117822261879286. doi:10.1177/1178222618792860
94 Areán PA, Pratap A, Hsin H, *et al*. Perceived Utility and Characterization of Personal Google Search Histories to Detect Data Patterns Proximal to a Suicide Attempt in Individuals Who Previously Attempted Suicide: Pilot Cohort Study. *J Med Internet Res* 2021;**23**:e27918. doi:10.2196/27918

95 Gomes de Andrade NN, Pawson D, Muriello D, *et al*. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philos Technol* 2018;**31**:669–84. doi:10.1007/s13347-018-0336-0
96 Lee N. Trouble on the radar. *Lancet* 2014;**384**:1917. doi:10.1016/S0140-6736(14)62267-4
97 Singer N. In Screening for Suicide Risk, Facebook Takes On Tricky Public Health Role - The New York Times. New York Times. 2018.https://nyti.ms/2RkXJCn (accessed 29 Sep 2020).
98 Barnett I, Torous J. Ethics, transparency, and public health at the intersection of innovation and Facebook's suicide prevention efforts. *Ann Intern Med* 2019;**170**:565–6. doi:10.7326/M19-0366
99 Horvitz E, Mulligan D. Data, privacy, and the greater good. *Science (80-)* 2015;**349**:253–5. doi:10.1126/science.aac4520
100 Devlin J, Chang M-W, Lee K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics 2019. 4171–86. doi:10.18653/v1/N19-1423
101 Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: Principles, methods and evaluation. *Egypt Informatics J* 2015;**16**:261–73. doi:10.1016/j.eij.2015.06.005
102 U.S. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS). https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm
103 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data 2016 31* 2016;**3**:1–9. doi:10.1038/sdata.2016.35
104 Lamberg L. Confidentiality and Privacy of Electronic Medical Records. *JAMA* 2001;**285**:3075–6. doi:10.1001/JAMA.285.24.3075
105 Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Rev* 2009;**57**:1701.https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006 (accessed 12 Nov 2018).
106 Shi X, Liu Q, Fan W, *et al*. Transfer learning on heterogenous feature spaces via spectral transformation. *Proc - IEEE Int Conf Data Mining, ICDM* 2010;:1049–54. doi:10.1109/ICDM.2010.65
107 Harel M, Mannor S. Learning from multiple outlooks. *Proc 28th Int Conf Mach Learn ICML 2011* 2011;:401–8.
108 Wei B, Paí CP. Cross Lingual Adaptation: An Experiment on Sentiment Classifications. 2010;:258–62.https://aclanthology.org/P10-2048 (accessed 5 Dec 2022).
109 Lee J, Yoon W, Kim S, *et al*. Data and text mining BioBERT: a pre-trained biomedical language representation model for biomedical text mining. doi:10.1093/bioinformatics/btz682
110 Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Published Online First: 10 April 2019.https://github.com/kexinhuang12345/clinicalBERT
111 Ji S, Zhang T, Ansari L, *et al*. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. Published Online First: 29 October 2021.https://github.com/Inusette/
112 Desautels T, Calvert J, Hoffman J, *et al*. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med informatics* 2016;**4**:e28. doi:10.2196/medinform.5909
113 Hwang AS, Atlas SJ, Cronin P, *et al*. Appointment "no-shows" are an independent predictor of subsequent quality of care and resource utilization outcomes. *J Gen Intern*

Med 2015;**30**:1426–33. doi:10.1007/s11606-015-3252-3
114 Avati A, Jung K, Harman S, *et al.* Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;**18**. doi:10.1186/s12911-018-0677-8
115 Simon GE, Johnson E, Lawrence JM, *et al.* Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry* 2018;**175**:951–60. doi:10.1176/appi.ajp.2018.17101167
116 Zheng L, Wang O, Hao S, *et al.* Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Transl Psychiatry* 2020;**10**. doi:10.1038/s41398-020-0684-2
117 Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clin Psychol Sci* 2017;**5**:457–69. doi:10.1177/2167702617691560
118 Wilimitis D, Turer RW, Ripperger M, *et al.* Integration of Face-to-Face Screening with Real-time Machine Learning to Predict Risk of Suicide among Adults. *JAMA Netw Open* 2022;**5**:E2212095. doi:10.1001/jamanetworkopen.2022.12095
119 McCoy TH, Castro VM, Roberson AM, *et al.* Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016;**73**:1064–71. doi:10.1001/jamapsychiatry.2016.2172
120 Zhang Y, Zhang OR, Li R, *et al.* Psychiatric stressor recognition from clinical notes to reveal association with suicide. *Health Informatics J* 2019;**25**:1846–62. doi:10.1177/1460458218796598
121 Pennebaker JW, Booth RJ, Francis ME. Linguistic Inquiry and Word Count: LIWC. 2007.http://www.liwc.net
122 Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J Lang Soc Psychol* 2010;**29**:24–54. doi:10.1177/0261927X09351676
123 De Choudhury M, Gamon M, Counts S, *et al.* Predicting Depression via Social Media. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. 2013. 128–37.
124 Huang X, Zhang L, Chiu D, *et al.* Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. In: *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*. IEEE 2014. 844–9. doi:10.1109/UIC-ATC-ScalCom.2014.48
125 Zirikly A, Resnik P, Uzuner ¨ Ozlem, *et al.* CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. *Proc Sixth Work Comput Linguist Clin Psychol* 2019;:24–33.https://www.
126 Shing H-C, Nair S, Zirikly A, *et al.* Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2018. 25–36. doi:10.18653/v1/W18-0603
127 MacAvaney S, Mittu A, Coppersmith G, *et al.* Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task. 2021;:70–80. doi:10.18653/v1/2021.clpsych-1.7
128 Gamoran A, Kaplan Y, Orr RI, *et al.* Using Psychologically-Informed Priors for Suicide Prediction in the CLPsych 2021 Shared Task. *Comput Linguist Clin Psychol Improv Access, CLPsych 2021 - Proc 7th Work conjunction with NAACL 2021* 2021;:103–9. doi:10.18653/v1/2021.clpsych-1.12
129 Eichstaedt JC, Smith RJ, Merchant RM, *et al.* Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A* 2018;**115**:11203–8.

doi:10.1073/pnas.1802331115
130 Corcoran CM, Cecchi GA. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2020;**5**:770–9. doi:10.1016/j.bpsc.2020.06.004
131 Elvevåg B, Foltz PW, Weinberger DR, *et al*. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr Res* 2007;**93**:304–16. doi:10.1016/J.SCHRES.2007.03.001
132 Gooding DC, Ott SL, Roberts SA, *et al*. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: Findings from the New York High-Risk Project. *Psychol Med* 2013;**43**:1003–12. doi:10.1017/S0033291712001791
133 Sonnenschein AR, Hofmann SG, Ziegelmayer T, *et al*. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cogn Behav Ther* 2018;**47**:315–27. doi:10.1080/16506073.2017.1419505
134 Cantor J, McBain RK, Kofner A, *et al*. Telehealth Adoption by Mental Health and Substance Use Disorder Treatment Facilities in the COVID-19 Pandemic. *Psychiatr Serv* 2022;**73**:411–7. doi:10.1176/appi.ps.202100191
135 Burkhardt HA, Alexopoulos GS, Pullmann MD, *et al*. Behavioral Activation and Depression Symptomatology: Longitudinal Assessment of Linguistic Indicators in Text-Based Therapy Sessions. *J Med Internet Res* 2021;**23**:e28244. doi:10.2196/28244
136 Cuijpers P, van Straten A, Warmerdam L. Behavioral activation treatments of depression: A meta-analysis. *Clin Psychol Rev* 2007;**27**:318–26. doi:10.1016/j.cpr.2006.11.001
137 Burkhardt H, Pullmann M, Hull T, *et al*. Comparing emotion feature extraction approaches for predicting depression and anxiety. In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2022. 105–15. doi:10.18653/v1/2022.clpsych-1.9
138 Sharma A, Miner AS, Atkins DC, *et al*. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. Published Online First: 17 September 2020.http://arxiv.org/abs/2009.08441
139 Sharma A, Lin IW, Miner AS, *et al*. *Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach*. Association for Computing Machinery 2021. doi:10.1145/3442381.3450097
140 Sharma A, Lin IW, Miner AS, *et al*. Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support. Published Online First: 2022.http://arxiv.org/abs/2203.15144
141 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:1–21. doi:10.1371/journal.pone.0118432
142 Shing H-C, Resnik P, Oard D. A Prioritization Model for Suicidality Risk Assessment. 2020;:8124–37. doi:10.18653/v1/2020.acl-main.723
143 Smucker MD, Clarke CLA. Time-Based Calibration of Effectiveness Measures. 2012;:95–104.
144 Walker J, Pan E, Johnston D, *et al*. The value of health care information exchange and interoperability. *Health Aff (Millwood)* 2005;**Suppl Web**:10–8. doi:10.1377/hlthaff.w5.10
145 Department of Health and Human Services. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. 2020. https://www.federalregister.gov/d/2020-07419/
146 Centers for Medicare & Medicaid Services. Medicare and Medicaid Programs; Patient Protection and Affordable Care Act; Interoperability and Patient Access for Medicare Advantage Organization and Medicaid Managed Care Plans, State Medicaid Agencies,

CHIP Agencies and CHIP Managed Care Entities, Iss. 2020. https://www.federalregister.gov/d/2020-05050/
147 Sayeed R, Gottlieb D, Mandl KD. SMART Markers: collecting patient-generated health data as a standardized property of health information technology. *npj Digit Med* 2020;**3**:1–8. doi:10.1038/s41746-020-0218-6
148 Mandl KD, Gottlieb D, Ellis A. Beyond One-Off Integrations: A Commercial, Substitutable, Reusable, Standards-Based, Electronic Health Record–Connected App. *J Med Internet Res 2019;21(2)e12902 https//www.jmir.org/2019/2/e12902* 2019;**21**:e12902. doi:10.2196/12902
149 HAPI FHIR - The Open Source FHIR API for Java. https://hapifhir.io/ (accessed 9 Jun 2020).
150 Bernert RA, Hilberg AM, Melia R, *et al*. Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *Int J Environ Res Public Health* 2020;**17**:1–25. doi:10.3390/ijerph17165929
151 Maguire M. Methods to support human-centred design. *Int J Hum Comput Stud* 2001;**55**:587–634. doi:10.1006/ijhc.2001.0503
152 Guest G, Bunce A, Johnson L. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field methods* 2006;**18**:59–82. doi:10.1177/1525822X05279903
153 Ancker JS, Benda NC, Reddy M, *et al*. Guidance for publishing qualitative research in informatics. *J Am Med Informatics Assoc* 2021;**00**:1–6. doi:10.1093/jamia/ocab195
154 Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;**15**:1277–88. doi:10.1177/1049732305276687
155 Saldaña J. *The Coding Manual for Qualitative Researchers (4th edition)*. 4th ed. Sage Publications, Inc. 2021.
156 Abraham J, Kannampallil T, Patel VL. A systematic review of the literature on the evaluation of handoff tools: Implications for research and practice. *J Am Med Informatics Assoc* 2014;**21**:154–62. doi:10.1136/amiajnl-2012-001351
157 Payne TH, Corley S, Cullen TA, *et al.* Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *J Am Med Informatics Assoc* 2015;**22**:1102–10. doi:10.1093/jamia/ocv066
158 Adler-Milstein J, Jha AK. Health information exchange among U.S. hospitals: Who's in, who's out, and why? *Healthcare* 2014;**2**:26–32. doi:10.1016/j.hjdsi.2013.12.005
159 Strasberg HR, Rhodes B, Del Fiol G, *et al.* Contemporary clinical decision support standards using Health Level Seven International Fast Healthcare Interoperability Resources. *J Am Med Informatics Assoc* 2021;**00**:1–11. doi:10.1093/jamia/ocab070
160 Coppersmith G. Quantifying Suicidal Ideation via Language Usage on Social Media. *Acta Anaesthesiol Scand* 2015;**49**:1387–90.https://qntfy.com/static/papers/jsm2015.pdf%0Ahttp://doi.wiley.com/10.1111/j.1399-6576.2005.00752.x%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/16146482
161 Low DM, Rumker L, Talkar T, *et al.* Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study. *J Med Internet Res* 2020;**22**:1–16. doi:10.2196/22635
162 Bates DW, Kuperman GJ, Wang S, *et al*. Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality. *J Am Med Informatics Assoc* 2003;**10**:523–30. doi:10.1197/jamia.M1370
163 Scott K, Lewis CC. Using measurement-based care to enhance any treatment. *Cogn Behav Pract* 2015;**22**:49–59. doi:10.1016/j.cbpra.2014.01.010
164 Lambert MJ. Is It Time for Clinicians to Routinely Track Patient Outcome? A Meta-Analysis. *Clin Psychol Sci Pract* 2003;**10**:288–301. doi:10.1093/clipsy/bpg025
165 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *J Gen Intern Med* 2001;**16**:606–13.

doi:10.1046/j.1525-1497.2001.016009606.x
166 Choi BCK, Pak AWP. A catalog of biases in questionnaires. *Prev Chronic Dis* 2005;**2**:A13.http://www.ncbi.nlm.nih.gov/pubmed/15670466
167 Mazzucchelli T, Kane R, Rees C. Behavioral Activation Treatments for Depression in Adults: A Meta-analysis and Review. *Clin Psychol Sci Pract* 2009;**16**:383–411. doi:10.1111/j.1468-2850.2009.01178.x
168 Alexopoulos GS, Arean P. A model for streamlining psychotherapy in the RDoC era: the example of 'Engage'. *Mol Psychiatry* 2014;**19**:14–9. doi:10.1038/mp.2013.150
169 Dillon DG, Rosso IM, Pechtel P, *et al*. Peril and pleasure: An RDoC-inspired examination of threat responses and reward processing in anxiety and depression. *Depress Anxiety* 2014;**31**:233–49. doi:10.1002/da.22202
170 Lewinsohn PM, Graf M. Pleasant activities and depression. *J Consult Clin Psychol* 1973;**41**:261–8. doi:10.1037/h0035142
171 Alexopoulos GS, Raue PJ, Banerjee S, *et al.* Comparing the streamlined psychotherapy "Engage" with problem-solving therapy in late-life major depression. A randomized clinical trial. *Mol Psychiatry* Published Online First: 2020. doi:10.1038/s41380-020-0832-3
172 Rude SS, Gortner EM, Pennebaker JW. Language use of depressed and depression-vulnerable college students. *Cogn Emot* 2004;**18**:1121–33. doi:10.1080/02699930441000030
173 Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. *J Res Pers* 2017;**68**:63–8. doi:10.1016/j.jrp.2017.02.005
174 Pyszczynski T, Greenberg J. Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychol Bull* 1987;**102**:122–38. doi:10.1037/0033-2909.102.1.122
175 Schwarzbach M, Luppa M, Sikorski C, *et al*. The relationship between social integration and depression in non-demented primary care patients aged 75 years and older. *J Affect Disord* 2013;**145**:172–8. doi:10.1016/j.jad.2012.07.025
176 Berkman LF, Glass T, Brissette I, *et al*. From social integration to health: Durkheim in the new millennium. *Soc Sci Med* 2000;**51**:843–57. doi:10.1016/S0277-9536(00)00065-4
177 Wiltsey Stirman S, Pennebaker JW. Word Use in the Poetry of Suicidal and Nonsuicidal Poets. *Psychosom Med* 2001;**63**:517–22. doi:10.1097/00006842-200107000-00001
178 Beck AT. *Depression: clinical, experimental, and theoretical aspects*. New York: : Hoeber Medical Division, Harper & Row 1967. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2237012/
179 Fast E, Chen B, Bernstein M. Empath: Understanding Topic Signals in Large-Scale Text. *Proc 2016 CHI Conf Hum Factors Comput Syst* 2016;:4647–57. doi:10.1145/2858036.2858535
180 Roberts K. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP. *Proc Clin Nat Lang Process Work* 2016;:54–63.
181 Kanter JW, Mulick PS, Busch AM, *et al*. The Behavioral Activation for Depression Scale (BADS): Psychometric Properties and Factor Structure. *J Psychopathol Behav Assess* 2007;**29**:191–202. doi:10.1007/s10862-006-9038-5
182 Cohen T, Widdows D. Empirical distributional semantics: Methods and biomedical applications. *J Biomed Inform* 2009;**42**:390–405. doi:10.1016/j.jbi.2009.02.002
183 Turney PD, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics. *J Artif Intell Res* 2010;**37**:141–88. doi:10.1613/jair.2934
184 Widdows D, Ferraro K. Semantic Vectors: A scalable open source package and online technology management application. 2008.
185 Widdows D, Cohen T. The Semantic Vectors package : New algorithms and public tools for distributional semantics. 2010.

186 Cohen T, Widdows D. semanticvectors/semanticvectors. https://github.com/semanticvectors/semanticvectors (accessed 12 Mar 2019).
187 Mikolov T, Sutskever I, Chen K, *et al*. Distributed Representations of Words and Phrases and their Compositionality. 2013;:3111–9.http://arxiv.org/abs/1310.4546 (accessed 12 Aug 2020).
188 Mikolov T. word2vec. 2013.https://github.com/tmikolov/word2vec
189 Lynskey D. Leonard Cohen: 'All I've got to put in a song is my own experience'. *Guard* 2012.https://www.theguardian.com/music/2012/jan/19/leonard-cohen (accessed 28 Jan 2021).
190 Hull TD, Malgaroli M, Connolly PS, *et al*. Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. *BMC Psychiatry* 2020;**20**:297. doi:10.1186/s12888-020-02721-x
191 Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: *9th Python in Science Conference*. 2010.
192 Lewinsohn PM, Sullivan JM, Grosscup SJ. Changing reinforcing events: An approach to the treatment of depression. *Psychother Theory, Res Pract* 1980;**17**:322–34. doi:10.1037/h0085929
193 Manos RC, Kanter JW, Luo W. The Behavioral Activation for Depression Scale-Short Form: Development and Validation. *Behav Ther* 2011;**42**:726–39. doi:10.1016/j.beth.2011.04.004
194 MacPhillamy DJ, Lewinsohn PM. The pleasant events schedule: Studies on reliability, validity, and scale intercorrelation. *J Consult Clin Psychol* 1982;**50**:363–80. doi:10.1037/0022-006X.50.3.363
195 De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*. New York, New York, USA: : ACM Press 2013. 47–56. doi:10.1145/2464464.2464480
196 Molendijk ML, Bamelis L, van Emmerik AAP, *et al*. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behav Res Ther* 2010;**48**:44–51. doi:10.1016/j.brat.2009.09.007
197 Chester A, Glass CA. Online counselling: a descriptive analysis of therapy services on the Internet. *Br J Guid Counc* 2006;**34**:145–60. doi:10.1080/03069880600583170
198 Titov N, Dear BF, Staples LG, *et al*. The first 30 months of the MindSpot Clinic: Evaluation of a national e-mental health service against project objectives. *Aust New Zeal J Psychiatry* 2017;**51**:1227–39. doi:10.1177/0004867416671598
199 Sagar-Ouriaghli I, Godfrey E, Bridge L, *et al*. Improving Mental Health Service Utilization Among Men: A Systematic Review and Synthesis of Behavior Change Techniques Within Interventions Targeting Help-Seeking. *Am J Mens Health* 2019;**13**:155798831985700. doi:10.1177/1557988319857009
200 Abramson LY, Seligman ME, Teasdale JD. Learned helplessness in humans: Critique and reformulation. *J Abnorm Psychol* 1978;**87**:49–74. doi:10.1037/0021-843X.87.1.49
201 Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc Natl Acad Sci U S A* 2017;**114**:E7900–9. doi:10.1073/pnas.1702247114
202 Zablotsky B, Terlizzi EP. Mental Health Treatment Among Adults: United States, 2019. *NCHS Data Brief* 2020;:1–8.
203 Rottenberg J. Emotions in Depression: What Do We Really Know? *Annu Rev Clin Psychol* 2017;**13**:241–63. doi:10.1146/annurev-clinpsy-032816-045252
204 Amstadter A. Emotion regulation and anxiety disorders. *J Anxiety Disord* 2008;**22**:211–21. doi:10.1016/j.janxdis.2007.02.004
205 Coppersmith G, Dredze M, Harman C. Quantifying Mental Health Signals in Twitter. In:

Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Stroudsburg, PA, USA, PA, USA: : Association for Computational Linguistics 2014. 51–60. doi:10.3115/v1/W14-3207
206 De Choudhury M, Counts S, Horvitz EJ, *et al*. Characterizing and predicting postpartum depression from shared facebook data. *Proc ACM Conf Comput Support Coop Work CSCW* 2014;:625–37. doi:10.1145/2531602.2531675
207 Kleinberg B, Vegt I van der, Mozes M. Measuring Emotions in the COVID-19 Real World Worry Dataset. In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics 2020. https://aclanthology.org/2020.nlpcovid19-acl.11 (accessed 12 Nov 2022).
208 Shen JH, Rudzicz F. Detecting Anxiety through Reddit. In: *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology —- From Linguistic Signal to Clinical Reality*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2017. 58–65. doi:10.18653/v1/W17-3107
209 Ekman P. Are There Basic Emotions? *Psychol Rev* 1992;**99**:550–3. doi:10.1037/0033-295X.99.3.550
210 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
211 Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: *NeurIPS Proceedings*. 2017. doi:https://doi.org/10.48550/arXiv.1705.07874
212 Saeb S, Lonini L, Jayaraman A, *et al*. The need to approximate the use-case in clinical machine learning. *Gigascience* 2017;**6**:1–9. doi:10.1093/gigascience/gix019
213 Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Can Med Assoc J* 2012;**184**:E191–6. doi:10.1503/cmaj.110829
214 Spitzer RL, Kroenke K, Williams JBW, *et al*. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch Intern Med* 2006;**166**:1092–7. doi:10.1001/archinte.166.10.1092
215 Plummer F, Manea L, Trepel D, *et al*. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *Gen Hosp Psychiatry* 2016;**39**:24–31. doi:10.1016/J.GENHOSPPSYCH.2015.11.005
216 Flemotomos N, Martinez VR, Chen Z, *et al*. Automated evaluation of psychotherapy skills using speech and language technologies. *Behav Res Methods* 2021;:690–711. doi:10.3758/s13428-021-01623-4
217 Liu T, Meyerhoff J, Eichstaedt JC, *et al*. The relationship between text message sentiment and self-reported depression. *J Affect Disord* 2022;**302**:7–14. doi:10.1016/j.jad.2021.12.048
218 Aoyama M, Sakaguchi Y, Morita T, *et al*. Factors associated with possible complicated grief and major depressive disorders. *Psychooncology* 2018;**27**:915–21. doi:10.1002/pon.4610
219 Ille R, Schöggl H, Kapfhammer HP, *et al*. Self-disgust in mental disorders - Symptom-related or disorder-specific? *Compr Psychiatry* 2014;**55**:938–43. doi:10.1016/j.comppsych.2013.12.020
220 Duggan M, Smith A. 6% of Online Adults are reddit Users. *Pew Res Cent* 2013;:1–10.https://www.pewresearch.org/internet/2013/07/03/6-of-online-adults-are-reddit-users/
221 Burkhardt HA, Laine M, Kerbrat A, *et al*. Identifying opportunities for informatics-supported suicide prevention: the case of Caring Contacts. In: *AMIA Annu Symp Proc*. 2022.
222 Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst* 2009;**24**:8–12. doi:10.1109/MIS.2009.36

223 Gururangan S, Marasović A, Swayamdipta S, *et al*. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. 2020;:8342–60. doi:10.18653/v1/2020.acl-main.740
224 Weiss K, Khoshgoftaar TM, Wang DD. *A survey of transfer learning*. Springer International Publishing 2016. doi:10.1186/s40537-016-0043-6
225 Howard J, Ruder S. Universal language model fine-tuning for text classification. *ACL 2018 - 56th Annu Meet Assoc Comput Linguist Proc Conf (Long Pap* 2018;**1**:328–39. doi:10.18653/v1/p18-1031
226 Plana D, Shung DL, Grimshaw AA, *et al*. Randomized Clinical Trials of Machine Learning Interventions in Health Care. *JAMA Netw Open* 2022;**5**:e2233946. doi:10.1001/jamanetworkopen.2022.33946
227 Hernandez-Boussard T, Lundgren MP, Shah N. Conflicting information from the Food and Drug Administration: Missed opportunity to lead standards for safe and effective medical artificial intelligence solutions. *J Am Med Informatics Assoc* 2021;**28**:1353–5. doi:10.1093/jamia/ocab035
228 Barthel M, Stocking G, Holcomb J, *et al*. Nearly Eight-in-Ten Reddit Users Get News on the Site. 2016. www.pewresearch.org
229 Naseem U, Lee BC, Khushi M, *et al*. Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. Published Online First: 2022.http://arxiv.org/abs/2204.04521
230 Bird S, Loper E, Klein E. *Natural Language Processing with Python*. O'Reilly Media Inc. 2009.
231 Wolf T, Debut L, Sanh V, *et al*. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2019;:38–45. doi:10.18653/v1/2020.emnlp-demos.6
232 Resnik P, Garron A, Resnik R. Using topic modeling to improve prediction of neuroticism and depression in college students. *EMNLP 2013 - 2013 Conf Empir Methods Nat Lang Process Proc Conf* 2013;:1348–53.
233 Chancellor S, Birnbaum ML, Caine ED, *et al*. A taxonomy of ethical tensions in inferring mental health states from social media. *FAT* 2019 - Proc 2019 Conf Fairness, Accountability, Transpar* 2019;:79–88. doi:10.1145/3287560.3287587
234 Timeline of WHO's response to COVID-19. https://www.who.int/news-room/detail/29-06-2020-covidtimeline (accessed 1 Jul 2020).
235 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4. doi:10.1016/S1473-3099(20)30120-1
236 Mcclellan M, Gottlieb S, Mostashari F, *et al*. A national COVID-19 surveillance system: Achieving containment. 2020.
237 Andersen M. Mobile Technology and Home Broadband 2019 | Pew Research Center. 2019.https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/ (accessed 9 Jun 2020).
238 Mayor S. Covid-19: Researchers launch app to track spread of symptoms in the UK. *BMJ* 2020;**368**:m1263. doi:10.1136/bmj.m1263
239 Drew DA, Nguyen LH, Steves CJ, *et al*. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science (80-)* 2020;:eabc0473. doi:10.1126/science.abc0473
240 Menni C, Valdes AM, Freidin MB, *et al*. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* Published Online First: 11 May 2020. doi:10.1038/s41591-020-0916-2
241 FHIRPath (Normative Release). http://hl7.org/fhirpath/ (accessed 11 Aug 2020).
242 Coronavirus (COVID-19) - Apple and CDC. https://www.apple.com/covid19/ (accessed 17 Aug 2020).
243 Symptoms of Coronavirus | CDC. https://www.cdc.gov/coronavirus/2019-

ncov/symptoms-testing/symptoms.html (accessed 17 Aug 2020).
244 Bay J, Kek J, Tan A, *et al.* BlueTrace : A privacy-preserving protocol for community-driven contact tracing across borders. *Gov Technol Agency, Singapore* 2020;:9.https://bluetrace.io/static/bluetrace_whitepaper-938063656596c104632def383eb33b3c.pdf
245 Corona Warn App from SAP & Deutsche Telekom | SAP News Center. https://news.sap.com/2020/06/corona-warn-app-deutsche-telekom-sap/ (accessed 11 Aug 2020).
246 RESTful API - FHIR v4.0.1. https://www.hl7.org/fhir/http.html (accessed 11 Aug 2020).
247 Google Inc. Flutter - Beautiful native apps in record time. https://flutter.dev/ (accessed 9 Jun 2020).
248 Keycloak. https://www.keycloak.org/ (accessed 17 Aug 2020).
249 Localization Platform for Translating Digital Content | Transifex. https://www.transifex.com/ (accessed 17 Aug 2020).
250 HL7/fhirpath.js: Javascript implementation of FHIRPath. https://github.com/hl7/fhirpath.js/ (accessed 11 Aug 2020).
251 Mandl KD, Kohane IS. No Small Change for the Health Information Economy. *N Engl J Med* 2009;**360**:1278–81. doi:10.1056/NEJMp0900411
252 Campbell R. The five 'rights' of clinical decision support. *J AHIMA* 2013;**84**:42–7; quiz 48.http://www.ncbi.nlm.nih.gov/pubmed/24245088
253 Shaw M. Writing good software engineering research papers. In: *Proceedings of 25th International Conference on Software Engineering (ICSE'03)*. 2003. 726–36.
254 Garcia-Ceja E, Riegler M, Nordgreen T, *et al.* Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive Mob Comput* 2018;**51**:1–26. doi:10.1016/j.pmcj.2018.09.003
255 Genes N, Violante S, Cetrangol C, *et al.* From smartphone to EHR: a case report on integrating patient-generated health data. *npj Digit Med* 2018;**1**. doi:10.1038/s41746-018-0030-8
256 Chen C, Haddad D, Selsky J, *et al.* Making Sense of Mobile Health Data: An Open Architecture to Improve Individual- and Population-Level Health. *J Med Internet Res 2012;14(4)e112 https//www.jmir.org/2012/4/e112* 2012;**14**:e2152. doi:10.2196/JMIR.2152
257 Estrin D, Sim I. Open mHealth architecture: An engine for health care innovation. *Science (80-)* 2010;**330**:759–60. doi:10.1126/SCIENCE.1196187/ASSET/FE2BE0BA-4218-432F-959C-24813C217239/ASSETS/GRAPHIC/330_759_F1.JPEG
258 gt-health/OMH-on-FHIR: OMH on FHIR project. https://github.com/gt-health/OMH-on-FHIR (accessed 7 Dec 2022).
259 Vaidyam A, Halamka J, Torous J. Enabling Research and Clinical Use of Patient-Generated Health Data (the mindLAMP Platform): Digital Phenotyping Study. *JMIR mHealth uHealth* 2022;**10**:1–17. doi:10.2196/30557
260 Interoperability Standards Advisory (ISA). United States Core Data for Interoperability (USCDI). https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi (accessed 15 Nov 2022).
261 Accessing Health Records | Apple Developer Documentation. https://developer.apple.com/documentation/healthkit/samples/accessing_health_records (accessed 7 Dec 2022).
262 EHR giant Epic explains how it will bring Apple HealthKit data to doctors | VentureBeat. https://venturebeat.com/business/ehr-giant-epic-explains-how-it-will-bring-apple-healthkit-data-to-doctors/ (accessed 7 Dec 2022).
263 Bergquist T, Buie RW, Li K, *et al.* Heart on FHIR: Integrating Patient Generated Data into Clinical Care to Reduce 30 Day Heart Failure Readmissions (Extended Abstract).

AMIA . Annu Symp proceedings AMIA Symp 2017;**2017**:2269–73.
264 Bass M, Rosen KD, Gerend MA, *et al.* Development and feasibility of a Configurable Assessment Messaging Platform for Interventions (CAMPI). *Fam Syst Heal* 2021;**39**:19–28. doi:10.1037/fsh0000592
265 IHE ITI Technical Committee. IHE IT Infrastructure Technical Framework Supplement Mobile Alert Communication Management (mACM). 2019.
266 Hong N, Wen A, Shen F, *et al.* Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open* 2019;**2**:570–9. doi:10.1093/jamiaopen/ooz056
267 Grieve G. FHIR and confusion about the 80/20 rule. Heal. Intersect. 2014.http://www.healthintersections.com.au/?p=1924 (accessed 8 Dec 2022).
268 FHIR Overview - Architects. http://hl7.org/fhir/R4/overview-arch.html (accessed 8 Dec 2022).
269 Norman DA. Steps toward a cognitive engineering. In: *Proceedings of the 1982 conference on Human factors in computing systems - CHI '82*. New York, New York, USA: : ACM Press 1982. 378–82. doi:10.1145/800049.801815
270 Cohen TA, Patel VL, Shortliffe EH. Reflections and Projections. In: *Intelligent Systems in Medicine and Health: The Role of AI*. 2022. 539–51. doi:10.1007/978-3-031-09108-7_20
271 Kannampallil TG, Franklin A, Mishra R, *et al.* Understanding the nature of information seeking behavior in critical care: Implications for the design of health information technology. *Artif Intell Med* 2013;**57**:21–9. doi:10.1016/j.artmed.2012.10.002
272 Dalai V V., Khalid S, Gottipati D, *et al.* Evaluating the effects of cognitive support on psychiatric clinical comprehension. *Artif Intell Med* 2014;**62**:91–104. doi:10.1016/j.artmed.2014.08.002
273 Adler-Milstein J, Aggarwal N, Ahmed M, *et al.* Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis. *NAM Perspect* 2022;**22**. doi:10.31478/202209c
274 World Health Organization. International classification of diseases index. Tenth revision. 1992.
275 Spackman K. SNOMED RT and SNOMED-CT. Promise of an international clinical terminology. *MD Comput Comput Med Pract*;**17**:29.http://www.ncbi.nlm.nih.gov/pubmed/11189756
276 Huff SM, Rocha RA, McDonald CJ, *et al.* Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary. *J Am Med Informatics Assoc* 1998;**5**:276–92. doi:10.1136/jamia.1998.0050276
277 ISO/HL7 21731:2006(en): Health informatics — HL7 version 3 — Reference information model — Release 1. https://www.iso.org/obp/ui/#iso:std:iso-hl7:21731:ed-1:v2:en (accessed 16 Nov 2022).
278 Pryor TA, Hripcsak G. The arden syntax for medical logic modules. *Int J Clin Monit Comput* 1993;**10**:215–24. doi:10.1007/BF01133012
279 Clinical Quality Language. https://cql.hl7.org/01-introduction.html (accessed 16 Nov 2022).
280 CDS Hooks. https://cds-hooks.hl7.org/ (accessed 25 Feb 2021).

Supplemental materials for Chapter 8

Supplemental Table 1. Functional requirements for ISACC.

MVP: Minimum viable product. ML: Machine learning.

Requirement	Part of MVP	Data model needs	Design opp. #
1. Patient enrollment			
1.1. SMART-on-FHIR app launch with single existing patient	x	Patient record	3.1, 3.2
1.2. Ability to create a brand new individual patient record (existing fEMR functionality)	x	Patient record	
1.3. Patient record setup			
1.3.1. Enter intervention notes	x	Patient/intervention-level notes	
1.3.2. Read-only display of EHR patient data for reference	x	Patient record Other resources referring to the patient	3.2
1.3.3. Select preferred name to use for patient (from available names)		Patient name	3.2
1.3.4. Specify communication mode (SMS vs. email vs. post card) and address (phone number, email address, or physical address) for patient (only SMS supported at first)	x	Patient contact information	3.2
1.3.5. Post-condition: Patient record marked as enrolled	x	Enrollment status	3.2, 3.3
1.4. Message schedule setup			
1.4.1. Pull in system-level schedule template	x	System-level message schedule template Patient/intervention-level message schedule	1.2
1.4.2. Populate correct date for birthday message	x	Patient birthday	1.2, 3.2
1.4.3. Populate holiday messages for all holidays between first and last regularly scheduled message			1.2
1.4.4. Populate patient's name for name placeholders	x	Patient record	1.2, 3.2
1.4.5. Allow date/time and message editing	x	Message-level date and time	
1.4.6. Allow adding and removing messages	x	Patient/intervention-level message schedule	
1.5. Editing			
1.5.1. Allow retroactive changes to preferred patient name			
1.5.2. Allow retroactive changes to preferred communication mode			
1.5.3. Allow retroactive changes to message schedule			
1.5.4. Scrap message schedule and create a new one			
2. Patient monitoring (list view)			
2.1. Display list of patients	x	Patient record	3.1, 3.2
2.2. Display patient enrollment status	x	Enrollment status	3.1, 3.2
2.3. Display last, first, sex, DOB, record number for each patient on patient list	x	Patient record	
2.4. Flag patients with an unanswered high-priority reply	x	Message priority Message response status	1.6
2.5. Flag patients with any unanswered reply	x	Message response status	1.6
2.6. Flag patients with an unanswered reply older than 24 business hours (1 business day)	x	Message response status Message response time	1.6
2.7. Allow user-level setting of what days are work days			
2.8. Default sorting: high risk on top, then in order of reply received	x	Message priority Message response time	1.6

2.9. Unenroll patient		x	Enrollment status	
2.10. Remove patient record		x	Patient record	
3. Work planning				
3.1. Staff assignment of patient				3.5
3.2. Automatic assignment of follow-up tasks based on availability				3.5, 1.7
3.3. Follow-up task status tracking				3.6
4. Messaging service (async worker)				
4.1. SMART-on-FHIR backend service with all enrolled patients		x	Patient record Enrollment status	3.1, 3.2
4.2. Monitoring replies				
4.2.1. Listen for replies from external service such as Twilio SMS API (only SMS supported at first)		x		
4.2.2. Score replies w/ ML		x	Message priority	1.6
4.2.3. Assign themes/markers w/ ML				1.8
4.2.4. Create Communication record (to save reply to DB) incl. score and themes		x	Message record	3.3, 3.4
4.2.5. Issue alerts for high-priority replies (only Email supported at first)		x	Message priority Message response status	1.5
4.2.6. Issue alerts for any received replies (only Email supported at first)		x	Message response status	1.5
4.3. Sending messages				
4.3.1. Trigger message sending when scheduled message time occurs		x	Patient/intervention-level message schedule	1.3
4.3.2. Create Communication record if successful		x	Message record Message delivery status	3.3, 3.4
4.3.3. Trigger auto-generated response under certain circumstances (TBD)				1.4
5. Patient message response authoring tool				
5.1. SMART-on-FHIR app launch with single existing patient		x	Patient record	
5.2. Message entry field, send message via appropriate pathways (via API) on accept		x	Message record	
5.3. Create Communication record when message is sent successfully		x	Message record Message delivery status	
5.4. Adding current impression to communication as note				
5.5. Information display				
5.5.1. Display essential patient information (name, sex, dob, contact info (phone, email, emergency contact), identifier)		x	Patient record	2.1, 3.2
5.5.2. Other EHR data				
5.5.2.1. Demographics		x	Patient record	
5.5.2.2. Diagnoses		x	Diagnosis records	
5.5.2.3. Care team		x	Care team records	
5.5.2.4. Appointments				
5.5.2.5. Suicide safety pan				
5.5.2.6. most recent PHQ-9, CSS		x	Observation records	
5.5.3. Display historical patient PRO scores (PHQ-9, CSS)				2.1, 3.1
5.5.4. Display history of exchanged messages		x	Message records	2.2
5.5.5. Display CARING CONTACTS intake notes		x	Patient/intervention-level notes	2.2
5.5.6. Display risk level (urgency) for current/visible message(s)		x	Message priority	2.3
5.5.7. Display historical message risk scores		x	Message priority	2.3
5.5.8. Display themes for current/visible message(s)				2.3
5.5.9. Display historical message themes				2.3
5.5.10. Display configurable list of resources (each with name/identifier and resource text)				2.4
5.5.11. Allow choosing configurable follow-up message template				2.4
5.5.12. Display empathy score of the message being written				2.5

5.5.13. Show rewrite suggestions				2.5
6. File patient note to EHR (When response authoring workflow is being finished)				
6.1. Allow configuration of note template(s) and note filing frequencies, with reasonable defaults				
6.2. Allow different note templates for different events (w/ configuration)				
6.3. Generate note text		x	Patient record Enrollment status Patient/intervention-level message schedule Message records	1.9
6.4. Allow note editing before filing		x		
6.5. Display templated note for copy-paste		x		
6.6. Create FHIR resource for note		x	Progress Note record	3.3