Published: 17 March 2023
Online AM: 07 October 2022
Accepted: 25 August 2022
Revised: 27 May 2022
Received: 12 November 2021

RESEARCH-ARTICLE

Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment

ANJA THIEME, Microsoft Research Cambridge, Cambridge, U.K.

MARYANN HANRATTY, SilverCloud, Dublin, Ireland

MARIA LYONS, SilverCloud, Dublin, Ireland

JORGE E PALACIOS, SilverCloud, Dublin, Ireland

RITA FAIA MARQUES, Microsoft Research Cambridge, Cambridge, U.K.

CECILY MORRISON, Microsoft Research Cambridge, Cambridge, U.K.

View all

Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment

ANJA THIEME, Microsoft Research
MARYANN HANRATTY, MARIA LYONS, and JORGE PALACIOS, SilverCloud Health
RITA FAIA MARQUES and CECILY MORRISON, Microsoft Research
GAVIN DOHERTY, Trinity College Dublin

Recent advances in AI and machine learning (ML) promise significant transformations in the future delivery of healthcare. Despite a surge in research and development, few works have moved beyond demonstrations of technical feasibility and algorithmic performance. However, to realize many of the ambitious visions for how AI can contribute to clinical impact requires the closer design and study of AI tools or interventions within specific health and care contexts. This article outlines our collaborative, human-centered approach to developing an AI application that predicts treatment outcomes for patients who are receiving human-supported, internet-delivered Cognitive Behavioral Therapy (iCBT) for symptoms of depression and anxiety. Intersecting the fields of HCI, AI, and healthcare, we describe how we addressed the specific challenges of (1) *identifying clinically relevant AI applications*; and (2) *designing AI applications for sensitive use contexts* like mental health. Aiming to better assist the work practices of iCBT supporters, we share how learnings from an interview study with 15 iCBT supporters surfaced their practices and information needs and revealed new opportunities for the use of AI. Combined with insights from the clinical literature and technical feasibility constraints, this led to the development of two clinical outcome prediction models. To clarify their potential utility for use in practice, we conducted 13 design sessions with iCBT supporters that utilized interface mock-ups to concretize the AI output and derive additional design requirements. Our findings demonstrate how design choices can impact interpretations of the AI predictions as well as supporter motivation and sense of agency. We detail how this analysis and the design principles derived from it enabled the integration of the prediction models into a production interface. Reporting on identified risks of over-reliance on AI outputs and needs for balanced information assessment and preservation of a focus on individualized care, we discuss and reflect on what constitutes a responsible, human-centered approach to AI design in this healthcare context.

The research of Gavin Doherty is funded in part by Science Foundation Ireland Grant no. 13/RC/2106 to the Adapt Centre.
Author's addresses: A. Thieme, R. F. Marques, and C. Morrison, Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, UK; emails: anthie@microsoft.com, v-rmarques@microsoft.com, cecilym@microsoft.com; M. Hanratty and M. Lyons, SilverCloud Health, One Stephen Street Upper, Dublin 8, D08 DR9P, IRL; email: maryann.hanratty@silvercloudhealth.com, maria.lyons@silvercloudhealth.com; J. Palacios, SilverCloud Health, 124 City Road, London, EC1V 2NX, UK; email: jorge.palacios@silvercloudhealth.com; G. Doherty, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, IRL; email: gavin.doherty@scss.tcd.ie.

1073-0516/2023/03-ART27 $15.00
https://doi.org/10.1145/3564752

CCS Concepts: • **Computing methodologies → Machine learning • Human-centered computing → Human computer interaction (HCI)**;

Additional Key Words and Phrases: Human-centered AI, human-centered machine learning, mental health, machine learning, decision-support systems, human-AI partnership, real-world implementation of AI, user research, IxD, responsible AI, ethics

ACM Reference format:
Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 27 (March 2023), 50 pages.
https://doi.org/10.1145/3564752

1 INTRODUCTION

Significant advances in AI and machine learning (ML) have led to ambitious visions of how new systems can revolutionize healthcare [55]. Continuing trends in personal health monitoring using mobile apps and wearables [61], combined with information increasingly collected in electronic healthcare records (EHR), contribute to a wealth of personal health and behavioral data that can be leveraged for health assessment, monitoring, and treatment [21, 44, 69, 71]. This growth in digital health data alongside improvements to computing power and cloud storage has led to a surge in AI research and development. The ability of advanced algorithmic models to mine structured knowledge of extensive data to discover previously unrecognized patterns is opening up new routes for improving our understanding of human behaviors, and predicting or optimizing health outcomes [14, 56, 100, 113].

In the field of medicine, AI applications can be wide-ranging. They have been particularly successful for image-based diagnosis (for example in radiology [7, 76, 113]); are used to aid interpretations of the human genome [8]; to help discover behavioral or biomarkers for understanding disease states and (sub-)types [14, 88, 71]; to predict patient outcomes such as hospital length of stay, chance of readmission, or mortality [17, 85]; to support the selection and adaptation of (drug) treatments [62]; or to facilitate the documentation and coordination of healthcare work [80].

Recent years have further seen a rapid growth in studies that explore AI applications in the domain of mental health [21, 93, 100]. These works seek to leverage AI for "social good" by helping to address the significant personal and economic burden that is caused by mental illness worldwide [107]. Here, most innovation has occurred in the areas of mental health diagnosis, symptom or risk detection [29, 72, 73, 74]—predominantly from sensor and text data. Less explored are approaches to help improve treatment access and delivery, which include developments of a chatbot or conversation-based approaches [38, 51, 79], and AI models to aid personalized treatment decisions [24, 77].

As early-stage research and development, the majority of these works demonstrate the technical feasibility and performance of achieved algorithms [1, 22, 37, 98], mostly from pre-existing datasets. This often leaves AI development removed from its target users or its study and integration within everyday (mental) healthcare, thereby limiting opportunities for desired real-world clinical impact [46]. Realization of many of the ambitious visions for AI-enabled healthcare transformation requires a closer study of AI systems and tools within specific health and care contexts, to better understand design opportunities and their implications on patients, clinicians, and other healthcare providers [6, 9, 30, 52, 95, 109, 111]. Here, HCI research and human-centered design approaches can make important contributions to help ensure future AI interventions are clinically useful, ethical, and can find acceptance and successful adoption in practice. In this regard, our research follows an applied agenda, seeking to involve users in the design of AI-based healthcare

systems and addressing some of the sociotechnical challenges that are involved in what has been termed as the "last mile" towards achieving real-world implementation [6, 26, 68].

As such, the work presented in this articleis one of the first to adopt and frame the development of an AI application for healthcare providers within an HCI methodology. Intersecting the fields of HCI, AI, and healthcare, we address two key challenges: (1) *how to identify what kinds of AI outcomes to develop* that enable the discovery of patterns in, and translation of, often complex data into insights that have clinical relevance and fit with the specific needs and practices of patients, clinicians, and health services; and, in response, (2) *how to design AI applications for sensitive use contexts* like (mental) healthcare such that non-AI experts can appropriately interpret provided AI outputs, and effectively and responsibly action these within their clinical decision-making and care provision.

More specifically, this articledescribes our iterative, human-centered approach to designing an AI application that predicts if a patient, who receives internet-delivered Cognitive Behavioral Therapy (iCBT) for depression and anxiety, will achieve a reliable improvement (RI) in their mental health symptoms by the end of treatment. This work forms part of a three-year, multi-disciplinary research collaboration that comprises a diverse team of researchers and developers with backgrounds in ML, Clinical Psychology, HCI, Design, Engineering, and Data Compliance. Jointly, we investigate identified AI challenges in the context of SilverCloud Health,[1] an established iCBT platform for the treatment of depression, anxiety, and functional impairments [86]. The platform offers guided self-help to patients, who work through offered therapy content by themselves in their own time. To promote engagement and the benefits from treatment, each patient is supported throughout the program by a human supporter [91, 110], who regularly communicates with them via online messages or phone conversations. These supporters are a specially trained cohort, typically graduate psychologists with further training in low-intensity iCBT interventions. They act as a facilitator of the computerized intervention through which users can learn and apply mental health self-management skills, and are the target users of our AI model outputs.

There are a number of factors that make this particular healthcare set-up particularly suited and feasible for the application and actual integration of AI. Firstly, as an established digital health service, the technology itself is already used at scale (>1,000,000 user base), which includes the routine collection of data about patients' treatment interactions and outcomes. Secondly, this information is processed and presented within data dashboards (Figure 2) to facilitate reviews of patient progress by the human supporters. Since these supporters already use interactive data dashboards as part of their routine work to gain insights about their patients for guiding clinical decision-making and next steps, it is more straightforward to imagine and practically realize the near-future integration of AI within this data review configuration. This paves the way for research to focus its explorations on the potential added value of advanced data analytics and the design of appropriate practices for their use, rather than common implementation barriers.

Describing our collaborative approach and reflections on what constitutes an ethical and responsible human-centered approach to AI design in this (mental) healthcare space, the article makes two main contributions:

(1) We describe how we *identified meaningful use scenarios and development targets for AI in this context.* We outline how the findings of an interview study with 15 iCBT supporters enabled an in-depth understanding of the information needs and data practices of our target users and revealed six areas of opportunity for AI. These learnings were reviewed by the research team and, together with feasibility and data availability constraints and

[1]SilverCloud Health: www.silvercloudhealth.com.

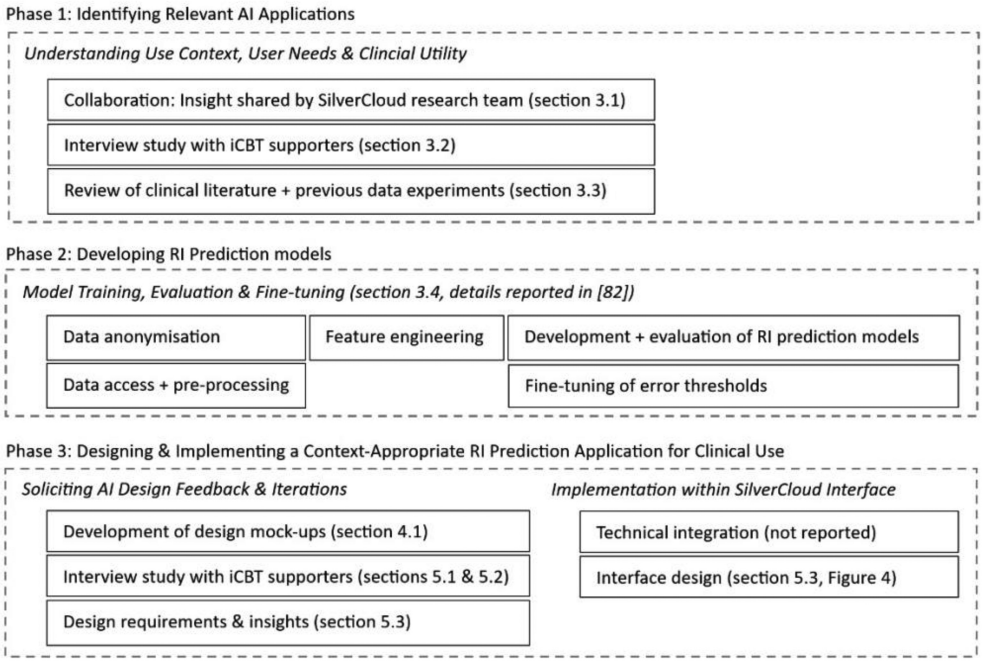


Fig. 1. Study overview.

insights from the clinical literature (feedback-informed therapy), determined our focus on predicting RI outcomes from iCBT program use. We outline the development of our final prediction models and reflect on key choices for their application in practice.

(2) We describe the *challenges for designing and integrating achieved predictions within an existing health service*. We report on 13 design sessions with iCBT supporters, which: (i) further clarified use scenarios for our AI models; (ii) demonstrated how design choices in representing the AI output shape perceptions of the purpose and interpretations of the prediction results; and (iii) revealed key concerns about the integration of AI-enabled data insights into existing support practices. These include demoralizing supporters or increasing performance pressures; uncritical treatment and over-reliance on AI predictions (especially by novice supporters); undesired changes to the intensity and nature of important "individualized" care; and implications of false predictions. Responding to these challenges, we describe how these learnings can translate into UI design and become integrated within the SilverCloud product. An overview of the research process and study steps is provided in Figure 1.

2 RELATED WORK

We begin with a brief overview of the current landscape of AI research in mental health, outlining in particular: (i) opportunities for developing AI applications that support treatment and personalized care; and (ii) the need for contextual integration of AI design and study to move forward developments of AI applications that are clinically useful, ethical, and that can find acceptance and adoption in real-world healthcare.

2.1 AI Applications in Mental Health: Open Opportunities for Treatment Support and Personalization

Recent years have seen a surge in research studying applications of AI in the domain of mental health (i.e., [21, 93, 100]). Where applications focus on affective mental health problems or conditions such as depression, anxiety, or stress—rather than neurological or neurodevelopmental disorders—existing works most commonly utilize (mobile phone) sensor data (i.e., [18, 29, 45, 74, 96, 106, 114]) or text, predominantly sourced from social media (i.e., [19, 58, 72, 78]), to *better understand; (earlier) detect; and (automatically) diagnose mental health status*. A smaller proportion of works describe AI approaches for *predicting mental health risks*; especially suicidality [1, 3, 73, 81, 102]. The predominant focus on AI for mental health detection and diagnosis can partially be explained by the challenges of requiring access to high-quality, large-scale mental health data. The costs involved in extensive data collection means that many existing studies use readily available data (e.g., public datasets, social media), or capture data from individuals described as "normal" or "healthy" (i.e., [29, 45, 74, 84, 98, 114]) rather than people with a diagnosed mental health condition. They then often apply standardized clinical scales or questionnaires to screen for specific mental health symptoms or their severity within their study population (i.e., [77, 84, 106]). Thus, general data access challenges combined with the availability of clinical outcome measures may explain the prevalence of algorithmic modeling for mental health symptom detection and diagnosis, which has shaped the kinds of research questions and algorithmic models that have been developed to date [100].

Beyond this trend, there are investigations into AI uses to help *scale up* or *improve the delivery of mental health treatment*. This includes the design of *conversation-based interfaces and chatbots* for either the delivery of psychotherapy [38, 51, 70, 108] or to support engagement with therapeutic activities [80] by enabling more open, honest self-disclosures [59, 66]. Prominent early examples include Woebot, a conversational agent for CBT delivery [38]; and Wysa, an emotionally-intelligent chatbot for users with self-reported symptoms of depression [51]. In addition to developing (fully automated) conversational agent experiences for delivering mental health and well-being interventions, AI techniques are employed to support text-based messaging with either a human coach or peer supporter. This includes the use of classifiers in the analysis of language markers to detect moments of "positive change" in the cognitive processes of people suffering from mental distress [83].

Within the context of treatment support, we find only a few examples of work that employ AI specifically to help *personalize interventions* to peoples' individual mental health and support needs [24, 77]. Here, Paredes et al. [77] developed a mobile phone app to recommend tailored coping strategies for stress management. To this end, their system learned from users' engagement with different stress interventions to predict which intervention—out of a given set—may be most correlated with stress reduction for a particular person. Moving from adaptations in content selection to personalized communication, Chikersal et al. [24] analyzed how specific linguistic features in support messages to patients receiving iCBT correlated with better patient outcomes dependent on the patients' specific circumstances (e.g., their current mental health, treatment week, level of engagement with iCBT). The research showed how certain aspects in the communication (e.g., use of positive words or words referencing social behaviors) correlated "more" or "less" with desired treatment outcomes for a particular patient context, enabling more tailored communications for those individuals.

In summary, there has been a surge in AI research and development to support predominantly mental health assessment; however, approaches to leveraging AI specifically to help improve digitally delivered psychotherapy interventions, such as iCBT, and support more personalized care, are currently under-explored.

2.2 Beyond Technical Performance: Towards Designing Real-world AI Applications for Mental Health

As identified in a recent review by Thieme et al. [100], the majority of current mental health work predominantly describes the technical development of (initial) algorithmic models as their main contribution alongside reports of specific methodological contributions (i.e., new approaches to data labeling [89, 112], or feature extraction [67]). Often positioned as proof-of-concept studies, these works tend to report the effectiveness of newly developed models based on their predictive performance; assessed via accuracy and error metrics (i.e., [1, 2, 3, 18, 22, 29, 32, 37, 41, 42, 43, 45, 58, 72, 73, 74, 75, 81, 84, 89, 96, 98, 112, 114]) and through comparison with other default or baseline models and state-of-the-art approaches (i.e., [1, 18, 22, 73, 74, 89, 96, 112, 114]). Yet, performance evaluations, typically based on held-out training data, may provide little insight into how reliably a model performs in the real-world; or how well-developed AI models could find useful adoption within existing healthcare practices (cf. [6, 26, 46, 95, 100]).

To date, most user involvement has been reported: in the collection of real-world user data (i.e., [2, 32, 75]); in data labelling with target-users or domain experts (e.g., [43, 73, 75, 112]); and for validating model results [36, 114]. Only a very small number of works present (participatory) design research [48, 52, 95, 103, 111] and user evaluation studies [49, 77] that: (i) deeply engage with the problems faced by potential users (i.e., patients, clinicians, mental health services) on a day-to-day basis; and (ii) extend understandings of how AI insights and applications could provide actual utility within those contexts. A key example here is the work by Hirsch et al. [48, 49] who conducted a participatory, iterative design process and pilot evaluation of an AI system that would automatically assess the motivational interviewing (MI) skills of psychotherapists from the audio of a face-to-face counseling session. It presents a rare example of research that explicitly engages with the design challenges of creating an interactive dashboard interface that presents the model outputs (derived from speech and language processing) in ways that are interpretable by humans, such that non-AI experts can develop an appropriate level of understanding and trust in AI model outputs. As such, the work considers how understanding of the AI emerges as part of real-world use, which expands much of existing HCI research on explainable AI (XAI i.e., [5, 13, 50]) that primarily contributes important methods and toolkits to address issues concerning the interpretability, fairness, accountability, and transparency of algorithms [6].

Additional examples can be found in the design of clinical decision support (CDS) systems [9, 15, 16, 52, 68, 95, 104, 111], mostly in clinical settings outside mental health. Yang et al. [111], for instance, describe an experience prototyping approach for integrating AI prognostics about the likelihood that a patient benefits from having an artificial heart implant. Positioning the AI in the corner of a "slide" that shows patient information for discussion within clinical, in-hospital meetings, they explored possibilities for more seamless AI output integrations within existing workflows. In other research, Cai and colleagues [15, 16] described user research and the development of a prototype prediction tool (called SMILY) for prostate cancer diagnosis. Their work surfaced varied information needs for onboarding medical practitioners in developing a human-AI partnership as part of clinical decision-making and demonstrated how the provision of interactive "refinement tools" increased the clinical utility of their tool as well as user trust in the algorithm. As part of a co-design process, Jacobs et al. [52] employed low-fidelity prototypes to study clinician's perceptions of different AI recommendations for what anti-depressant to administer to patients suffering from Major depressive disorder. What these and other works highlight is the gap in existing research and challenges for the successful use and implementation of AI systems within real-world clinical care; and the need for a closer design and study of AI applications within the social structures and (physical or digital) work environments that characterize the often complex eco-systems that surround healthcare provision (i.e., [9, 68]).

In summary, despite many important and innovative technical advances to date, more research and a closer integration of AI design within (mental) healthcare contexts is required. A deeper understanding and close response to stakeholder needs and expectations are necessary if we want to take steps forward in achieving AI applications that are clinically useful and that can find acceptance and adoption within routine clinical care.

3 UNDERSTANDING THE USE CONTEXT AND IDENTIFYING MEANINGFUL AI OPPORTUNITIES

This section addresses our first objective: *our approach to identifying useful applications and development targets for AI in our specific mental healthcare context.* We begin by describing how human-supported iCBT is delivered through SilverCloud. Reporting key findings of an interview study with iCBT supporters, we articulate our choices for the system we developed to (i) contribute to feedback informed therapy (FIT); (ii) predict RI outcomes; and (iii) better account for implications of different AI error types for its use in practice.

3.1 Background: CBT offering via SilverCloud Health and the Role of iCBT Supporters

The SilverCloud Health platform offers a wide range of self-guided iCBT programs whose clinical effectiveness has been evidenced through rigorous clinical research (e.g., [34, 86, 87]). Each program contains a set of core psycho-educational and psycho-therapeutic modules that are delivered using interactive, multi-modal contents [24, 31]. While iCBT is mostly self-administered, research has shown how the involvement of a human supporter or coach, who guides and assists the person, improves user engagement in therapy and leads to more effective mental health outcomes than unsupported interventions [24, 54, 91, 110]. Thus, although patients work through the program content at their own pace and time, they receive support from a trained supporter in the form of weekly or bi-weekly reviews throughout their treatment journey. Most supporters are graduate psychologists with further training in low-intensity interventions, including iCBT. Within the NHS England IAPT[2] initiative, they take on the role of a *Psychological Wellbeing Practitioner*[3] (PWP), who offers evidence-based interventions to patients with mild-to-moderate symptoms of depression and anxiety.

PWPs are well-trained in guiding patients on the use of the platform, recommending treatment content, and helping patients work through identified difficulties to both support good patient experiences and desired clinical outcomes. To this end, all interactions between the PWP and their patients are coordinated via an Intervention Management Site (Figure 2). Here, PWPs can review patient messages, their responses to clinical questionnaires; visits of therapy content (pages); completion of treatment tools; and the frequency of program logins. In response to this data, PWPs provide patients with feedback, typically by selecting and adapting an online messaging template within the intervention site. These templates are written by PWPs in their own words and tailored to each patient. Typically, a personalized patient review takes PWPs 10–15 minutes to complete.

While it can be more challenging to convey a "human touch" and nurture a therapeutic alliance through a computerized treatment format, to establish a person-centered relationship between supporter and patient is crucial in helping to (i) reduce perceptions of support being delivered through a machine rather than a human who listens, understands, and cares for the person; (ii) improve patient engagement and their belief in the effectiveness of an "online" approach; and (iii) ensure that the same standard and quality of care is delivered via an online treatment as is in other therapy formats; all of which enhance overall therapy outcomes from iCBT.

[2]NHS IAPT (Improving Access to Psychological Therapies): https://www.england.nhs.uk/mental-health/adults/iapt/.
[3]PWP role: https://www.instituteforapprenticeships.org/apprenticeship-standards/psychological-wellbeing-practitioner/.

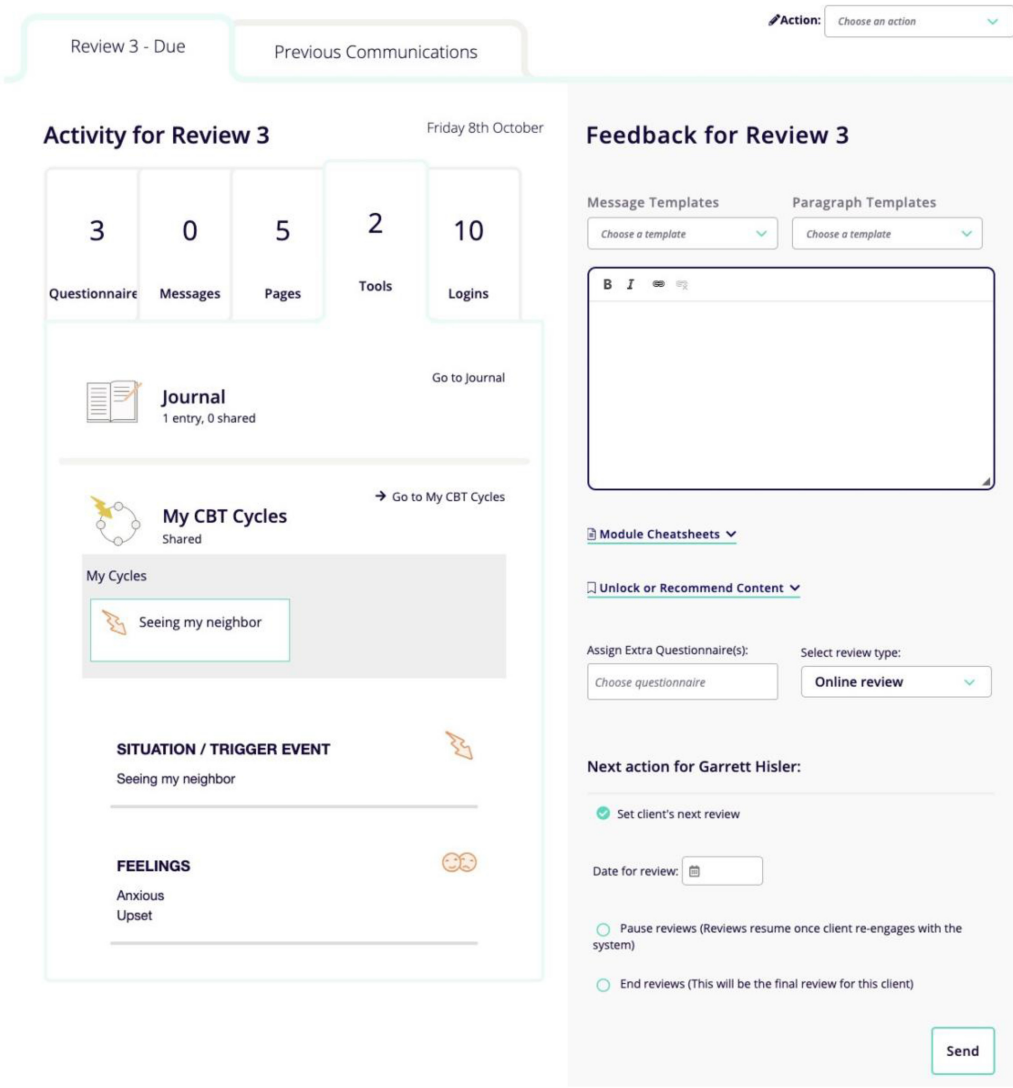


Fig. 2. Extract of an example patient review page within the supporter interface, showing: the frequency and details of patient completed clinical questionnaires; the number and content of messages sent to their supporter; therapy pages viewed; tools used; and overall system logins. Each of these summary items can be expanded for more detail. Following their review, supporters write a personalized message for which they can select and adapt text templates. They can also "bookmark" existing or "unlock" additional contents to advance user engagement and treatment progress.

Yet, the need to maintain timely, responsive patient care is complicated by generally high workloads and demands on mental health services. As the PWP role is often a career stepping stone for supporters, this is coupled with a moderate rate of staff turnover, which can result in high variations in the level of expertise of these PWPs, including many trainees and novices. This raises the question of how to best maximize the effects and outcomes of human support in this format and help iCBT supporters (of varying expertise) in effectively guiding their patients through treatment.

3.2 Understanding iCBT Supporter Work Practices and Identifying AI Opportunities

Seeking to open-up the design space and investigate how methods of AI could be leveraged to benefit supporter work practices, we conducted 15 semi-structured, 1 hourly interviews with PWPs to better understand their information needs and challenges in understanding patient progress and providing effective, personalized feedback. Responding to these learnings, we scoped out potential AI opportunities. Appendix A provides an overview of the study methods and a summary of the findings, including associated AI opportunities. A detailed report is available here [99]. Below, we restrict reports of key learnings to one of the identified themes: *understanding patient mental health: risks, progress and barriers* that specifically guided our AI development.

3.2.1 Study Theme "Understanding Patient Mental Health: Risks, Progress, and Barriers". A key focus in the review practices of PWPs is to gain a sufficient understanding of their patients' mental health. In the first instance, this includes (i) gaining an understanding of the person's *main mental health problem* to be able to *identify the right treatment*, and *effectively focus* or *adapt the treatment* to the person. To this end, supporters collate patient information from initial screening and treatment calls that are recorded in a *patient boarding card* within the care providers EHR system together with other relevant health information.

During regular patient reviews, supporters then assess: (ii) whether there are any *indicators of risk* that have to be responded to immediately, and (iii) how the patient is *progressing in their mental health* throughout treatment, as one of the first and most important steps in their review. Predominantly, this is done by examining changes in the patients' mental health symptoms as reported via standardized clinical questionnaires of PHQ-9 [60] (a 9-item measure of depression symptoms) and GAD-7 [64] (a 7-item measure of anxiety symptoms).

A "mental health risk" is typically indicated through a "risk tab" within the supporter interface that flags- when a patient scores higher on question nine of the PHQ-9, which asks about the person's thoughts on self-harm and intent to end their life. In some cases, mental health risk indicators are also picked up in patient's messages, or through changes in treatment engagement (i.e., absence of communication with the supporter, reduced program use). In cases of an identified risk, the supporter would often take action by calling the patient to assess what is going on for them and to clarify if their current scores actually reflect how they are feeling.

For assessing the "mental health progress" of a patient, PWPs review trends in patients' clinical scores. If those scores show a gradual improvement, little time is spent reviewing these further. However, if the scores remain unchanged; change suddenly or significantly in either direction; or indicate a decline in mental health, then the supporter would take further action. For example, if the patient scores are not improving or are not at the level that the supporter expected them to be, they will look for potentially "barriers". This can include strategies such as: (i) asking the patient in an online message or call how they are feeling to better understand their circumstances; or (ii) taking a closer look at key aspects of clinical questionnaire items on which the patient may score highly that could be indicative for example of "persistent sleep difficulties" or "high levels of anxiety".

In cases where a patient does not improve after several weeks, they are brought to case management supervision (CMS) to re-assess the suitability of the current treatment approach for them. In general, all patients are regularly reviewed in CMS, typically: when they are *new* to the service; *four weeks* into an intervention; had *no changes in their scores*; or showed *risks*. In CMS, supporters discuss the patients' main mental health problem and current treatment plan and consider their level of program engagement and clinical scores to identify *risks of drop-out* and strategies to address any *treatment struggles* (i.e., by changing the intervention, or stepping the person up to a more intensive care approach).

It is important to note that some supporters also described the need for caution in interpretations of the clinical scores, as they may not exactly reflect what is going on for the person. While frequent clinical assessments can provide a numerical indicator and trend of that person's mental health progression, these should not be given too much weight or consideration in isolation. Supporters, therefore, described efforts to *assess clinical scores in the context of other information* provided in patient messages and conversations. For example, additional information can help mental health assessment by providing contextual insights into why someone might be struggling with stress and worries (e.g., exam times for students), and can clarify if someone has easy access to the means to commit suicide (e.g., an isolated farmer) that require action.

3.2.2 Summary and Opportunities for AI. In summary, supporters often gain an important understanding of their patients' main mental health problem and life circumstances through an initial assessment call. Combined with program engagement reviews, these serve as contextual resources to evaluate patients' mental health progression and guide treatment. A key focus in assessing patient mental health progress involves the regular review of clinical scores and their changes over time to detect any *mental health risks*, potential *barriers to improvement,* and *requirements to re-evaluate the patients' suitability for a particular treatment* and *adapt their care*. Responding to these tasks, we identified the following (not fully exhaustive) list of questions and opportunities for AI that could be investigated further:

- How could AI support early or automatic detection of mental health risks?
- How could AI assist supporters' understanding of why a patient might not be improving in their mental health?
  - How could AI support the identification of "barriers to patient improvement" based on patients' clinical scores or specific markers in their language (i.e., in patient messages and text entries)?
  - How could AI be leveraged to learn what mental health struggles (i.e., indicated via questionnaire items) may be particularly salient in predicting good or poor mental health outcomes?
- How could AI help identify "early" if a patient is likely to benefit from the chosen treatment, or not, to facilitate treatment adjustments and ensure patients get the "right care"?

3.3 Rationale for a Focus on Clinical Outcome Prediction

The user research findings, including the theme of mental health outcomes outlined above, enabled important, in-depth understandings of the review practices and personalized treatment goals of the supporters. As a next step, our multi-disciplinary research team assessed these findings and associated proposals for AI applications in the context of (i) available data constraints, and (ii) the clinical literature to identify how AI-enabled insights or interventions could both be meaningful in supporting supporter practices, as well as (iii) practically feasible.

Identifying AI applications that usefully integrate needs for clinically relevant insights with data availability constraints and workflow implementation challenges: Amongst the data available to us are clinical measures of PHQ-9 and GAD-7 that are frequently completed as part of regular patient reviews as well as behavior engagement data that includes information such as what sections of the treatment program a patient used (i.e., content pages; interactive therapy tools; a journal; profile page; interactions with their supporter) over time. Prior to any analysis, all data is carefully de-identified through the removal of any private data such as free text entries, demographic information, and dates (see [82] for details); restricting any investigations to higher-level program use features to protect patient anonymity. Our analysis further focuses on SilverClouds' Space from

the Depression and Anxiety program, which has the largest number of patients enrolled to ensure a sufficiently large, representative dataset for the use of AI methods.

As a multi-disciplinary team, we explored numerous avenues for extracting meaningful data insights to assist supporters in delivering effective, personalized iCBT care. Amongst others, this included the development of a probabilistic latent variable model to better understand patterns of patient engagement with the iCBT program [23]; and a sequential decision-making model that suggests what therapy content (i.e., a specific iCBT module) at a particular time in treatment may best correlate with improved patient outcomes (i.e., symptom reduction) to guide supporter recommendations. Discussing these modeling efforts and the user research findings; our clinical collaborators described the potential of using AI in contributing to FIT.

The clinical literature on FIT indicates that offering therapists feedback on the expected outcomes that a patient might attain from treatment can improve its success and prevent symptom deterioration in patients who are not progressing well (e.g., [28, 39, 40]). This is of particular relevance since, in the case of depression and anxiety; just over half of all patients are expected to respond to treatment, and those that do remain at considerable risk of future relapse [4, 10]. Thus, the ability to determine at an early stage whether or not a patient is likely to improve can allow iCBT supporters to assess more closely whether treatment is progressing as expected, and to adapt its delivery if necessary, in order to increase the likelihood of improvement in patient symptoms (see [11, 21, 27, 47, 65] for earlier explorations in this area). In other words, providing insights early about *prospective patient outcomes* can create opportunities for supporters to engage in more timely and proactive interventions to ensure treatment success. Such insights can also usefully guide decision-making within clinical management supervision (CMS) to aid decisions, i.e., about whether a patient needs to be stepped up to different care sooner, or whether more treatment sessions would need to be administered to a particular patient. All this may also assist in more effective care planning through an earlier re-distribution of care resources, and reduce delays and negative effects of having a patient attend for too long the wrong care pathway. This motivated the decision to develop a state-of-the-art prediction model that can provide supporters, early into treatment, with important insights about the likelihood that their patients will achieve desired mental health outcomes by the end of treatment.

More specifically, we chose to predict the likelihood of *RI* in mental health symptoms as the desired mental health outcome as it reflects a significant positive change in patient symptoms and presents a core performance metric used by NHS IAPT services in the UK to determine treatment success ([25], p.154). RI is defined as a decrease in PHQ-9 score of 6 or more points; and a decrease in GAD-7 score of 4 or more points, at the end of treatment. This outcome metric serves as an indicator of real improvement in symptoms, as it exceeds change that can be accounted for by measurement error and is generally reported to capture how many people showed any real benefit from the treatment delivered [53].

3.4 Overview on Developed Outcome Prediction Models for RI in Depression and Anxiety Symptoms

To predict RI in patient symptoms of depression and anxiety, we analyzed fully anonymized data of 46,313 patients who: were enrolled in the Space from Depression and Anxiety program between January 2015 and March 2019; had a supporter assigned to them; and completed clinical symptom measures of PHQ or GAD at least twice. For these patients, we used time-stamped scores from the completion of clinical measures—their PHQ-9 and GAD-7 scores—as the inputs (so called data *features*) to train two recurrent neural network (RNN)-based models (see Appendix B for an illustration of the RNN architecture). Based on the validation dataset, these RNNs achieve an overall accuracy of 83.75% in predicting RI in PHQ-9 outcomes; and 78.86% in

predicting RI in GAD-7 outcomes; and consistently outperform other baseline models (LogR, RF, GBMs, EMA). See [82] for more details. The models predictive performance improves with time, meaning that with three or more clinical measures available[4], the RNNs achieve accuracies of 87.77% for PHQ-9 (Specificity = 95%, Sensitivity = 66.1%) and 87.37% for GAD-7 (Specificity = 95%, Sensitivity = 70.27%). This means that both RNNs achieve above 87% prediction accuracy with high specificity (set to 95%) and reasonable sensitivity after three review periods during which the patient completes clinical measures. In other words, rather than just considering the overall accuracy of our model (across the entire receiver operating curve) [26], we chose to fine-tune our algorithm towards higher specificity, which is reflective of a low *false positive* rate.

Weighing up which error type is most costly to define appropriate AI performance thresholds: For our specific application context, *false positive* errors would mean that a patient, who would need extra help would be at risk of not receiving it (due to a prediction that falsely indicates that the person will achieve RI when they do not). Contrary, if the model would falsely predicts patients to not achieve RI when in fact they do (high sensitivity meaning low *false negative* rate), this might have less negative implications for patients and care providers alike as the desired outcome is achieved in the end. However, false negative errors can nonetheless be problematic as they can mean extra resources are allocated unnecessarily in supporting patient recovery, and they can cause disruptions to a patient's treatment journey if they were referred to more intensive care than needed. Our decision to set our model specificity to above 95% (low false positive rate), at the expense of a higher false negative rate, was also informed by clinician's assessment that for most negative RI predictions, whether they were true or false, they would work harder to identify the patient's difficulties and treatment needs. As such, they felt that false negative errors were likely not to weigh in as much as false positives (not receiving extra help when needed). This surfaced false positives as the main error type to avoid.

As described in other recent AI studies within clinical contexts [9, 16], it is important to carefully consider and weigh up how different error types may come to impact clinical decision-making and associated real-world implications in terms of costs and burdens on patients and health services. There is a need for future work to identify avenues (and potentially new tools) to assist AI developers and domain experts: (i) to learn more about the risks and costs that may occur through different prediction errors; (ii) to identify what constitutes an appropriate threshold between false positives and false negatives given the specific use and care context; and (iii) to re-evaluate those thresholds once their implications are better understood (i.e., through further investigations and use). Cai et al. [16] further suggest making such algorithmic design decisions (i.e., to optimize for a low false positive rate) transparent to users alongside other information such as the models' intended uses.

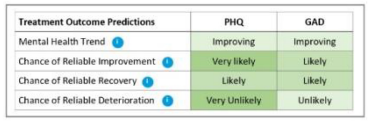
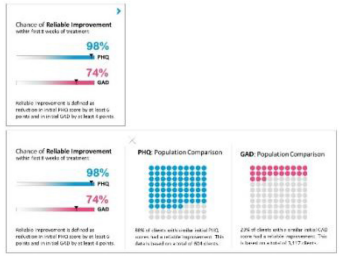
Having achieved prediction models with high accuracy (>87%) and few false positives after three reviews, were assessed by our clinical collaborators to be "good enough" to be considered for real-world deployment.

4 STUDY: UNDERSTANDING AND DESIGNING FOR APPROPRIATE USE OF OUTCOME PREDICTION

Having specified our focus on predicting patient outcomes, specifically RI in depression and anxiety symptoms, to support iCBT supporters work practice, this section turns to our second research

[4]Within the context of SilverCloud, PHQ-9 and GAD-7 questionnaires are administered by the supporter at the beginning of a "new review period". A review period is completed once the supporter sends a personalized feedback message as a response to reviewing patient progress, typically every 1-2 weeks during active treatment, which is mandatory for all NHS IAPT services (digital and non-digital). Post-receipt of this feedback message, patients are then prompted to complete the clinical questionnaires upon their first login when returning to the iCBT program.

Table 1. Shows the Five Design Mock-ups and Their Rationale for Discussing Different Ways in Which to Visualize and Otherwise Contextualize Outcome Prediction Results within the PWP Intervention Interface

Concept	Design	Rationale
1. Wheel Indicator		The design was inspired by a Geiger counter that acts as a strong visual in amplifying the binary prediction output (*yes* or *no* to Reliable Improvement). Conforming with common Western connotations for colors and positioning, the desirable outcome (*yes* to RI) is placed to the right and emphasized through green (in contrast with red). Model confidence is communicated through combining a color gradient in the wheel with the positioning of the counter; extended by a percentage value (i.e., 98% certainty).
2. Multiple Outcomes Table		The idea behind this design is to have a simple way of stating, through written text, whether the person was 'likely' or 'unlikely' to achieve RI. A combination of word categories (very unlikely, unlikely, uncertain, likely, and very likely) and color gradient (the darker the better) are chosen to communicate model confidence in its prediction. The table further includes other common outcome metrics such as 'reliable recovery', 'reliable deterioration' and 'mental health trend' to probe with participants into preferences for different metrics, and the potential relevance of having multiple mental health outcomes shown at once.
3. Visual Cue + Text		This design combines the use of textual categories with a simplified visual cue that presents as a small bar charting three regions to mark: negative prediction, model uncertainty, and positive prediction. A small arrow on top indicates the model output and its confidence through its relative position within the output region. In combining text categories with a visual cue, the aim is to reduce ambiguity in interpretations of the text labels without showing specific numerical values, which can be misleading at times (i.e., small changes in percentage values of model uncertainty can be difficult to make sense of).
4. Population Comparison		This design illustrates processes of layering of information and data comparisons. It is the only concept that shows a definition of RI upfront rather than requiring its retrieval by clicking on an information icon. The actual prediction is visualized as progress bars, indicating how far the patient has come towards the reliable improvement target for both PHQ and GAD; amplified through a percentage number. The representation can further be unfolded to receive additional information of how the prediction for the particular patient compares to other patients who have similar PHQ or GAD profiles; thereby potentially highlighting how representative (common or unusual) the predicted outcome is for this patient(type). To aid interpretation, a visual plot is added to show proportions alongside text that describes the data origin and its scale.
5. Dashboard Feedback		This design is proposed as an extension to the already existing supporter dashboard that displays to a supporter all of their SilverCloud patients within a comprehensive table. Amongst others, the table indicates the date a specific patient is due for a supporter review and how actively patients engage with treatment. It predominantly serves as an entry point for navigating to each individual patients' profile. The proposal is to add to it functionality whereby the prediction for RI is shown as percentage numbers in additional columns (for PHQ and GAD). To avoid risks of potentially demotivating supporters, a deliberate choice was made to communicate the prediction outcome positively, even if desired outcomes are unlikely to be achieved for some or all patients. To this end, the prediction output is translated into a spectrum of 0-100% to communicate progress towards the desirable target of RI, and all percentage numbers are shown in green.

objective: *how to design our AI application such that supporters can appropriately interpret provided data outputs, and effectively and responsibly action these within clinical decision-making and care.*

More specifically, we needed to clarify: (i) the specific use scenario and integration of the prediction output within existing workflows; (ii) the design of effective ways to communicate the prediction output within the supporter interface; and (iii) any foreseeable concerns or risks associated with the use of outcome prediction in this context. To this end, we conducted a series of design sessions with PWPs for which we developed interface mock-ups (Table 1) to concretize the AI outputs for PHQ-9 and GAD-7, and facilitate additional learnings about design requirements.

Next, we detail the study and its key findings, which brought forward a set of design sensitivities and requirements that cumulated into the implementation of a first UI design (Figure 4).

4.1 Development of Design Mock-Ups for Outcome Prediction

Over a period of six weeks, members of the research team (AT, RM, GD) regularly met with the UX team (MH, ML) at SilverCloud to develop a set of design mock-ups that would enable us to probe with study participants where, when and how outcome prediction results could add value to their work practices; what would constitute an easy-to-comprehend representation of the proposed AI output; and to learn about potential needs to add any contextual information to aid an appropriate level of understanding of probability-based prediction outputs.

Our conversations began with joint ideation sessions that culminated in the development of numerous design sketches. Partially, our design ideas were inspired by common statistical approaches that often use ranges and standard deviations to indicate variance in data results; as well as other AI design examples reported in the literature. This includes communications of prediction accuracy through percentages [52, 90]; illustrations of model uncertainty via diffuse color regions [57]; or temporal line charts of a patient health "prognosis" to illustrate trends over time [111]; as well as indications of risks (i.e., the chance to catch a specific disease relative to a standard population) through bar charts, line charts and other proportionate representations (i.e., [63, 97, 103]). Visual (and interactive) mechanisms to inspect model accuracies or prediction uncertainties are also commonly used by data scientists to better understand and debug their AI models. Amongst others, they demonstrate the importance of certain data features in explaining a prediction outcome, i.e., through (additive) bar charts [50], or communicate data saliencies through heat maps [23, 24]. As described previously, empirical research that studies how health or care professionals perceive and interpret AI outputs as part of user interfaces or within their work context however remains rare (exceptions include [35, 49, 52, 68, 95, 109, 111]).

To invite meaningful dialogue on use scenarios and design requirements with our iCBT supporters, we wanted to create a set of concrete design concepts that were distinct in a number of ways:

- *placement of prediction outcome within PWP interface* (single patient profile vs. all patients dashboard);
- *level of data abstraction* (individual patient vs. patient cohort);
- *modality* (graphical, numerical or textual representations);
- *complexity* (basic visual or single numeric value vs. more layered information); and
- *communication of model confidence* (single value vs. range as indicated through size or color gradients).

Varying across these dimensions, Table 1 illustrates the five design concepts that we selected for the study alongside descriptions of their design rationale. Concepts 1–4 are presented to participants within the context of a patients' individual review profile that shows their mental health symptom trajectory over time as exemplified in Figure 3; whereas concept 5 shows the prediction for all patients of a supporter in an overview dashboard.

4.2 Participants

We recruited 13 PWPs from the Berkshire Healthcare NHS Foundation Trust in the UK. Our sample includes seven participants who had taken part in our previous interview study; those repeat participants are indicated through grey shading in Table 2. Study participants predominantly self-reported as female (2 male) and included fully certified PWPs who were very experienced at using SilverCloud to support their patients. Asked to rate their own level of experience as a SilverCloud supporter, half of the participants (n = 7) self-identified as "intermediate/ expert" or "expert"; with

Fig. 3. As illustrated here through the Wheel indicator, design concepts 1–4 were shown within the context of a patient's individual review profile that shows their mental health symptom trajectory over time as derived from the completion of standardized clinical questionnaires of PHQ-9 and GAD-7.

Table 2. Participants' Role Title and Level of Expertise in Supporting Patients via SilverCloud; Amount of Years They Have Been Using the Service; and Number of Patients Currently Assigned to Them

Current Role	Supporter expertise	Years using SilverCloud	Patient load
PWP	Intermediate	0–2	15–30
PWP	Intermediate	0–2	15–30
PWP	Intermediate/ Expert	1–2	15–30
PWP	Intermediate	1–2	30 or more
PWP	Novice/Intermediate	2–4	30 or more
PWP	Intermediate/ Expert	2–4	15–30
PWP	Intermediate	4–5	5–15
PWP	Expert	2–4	30 or more
Senior PWP	Intermediate	2–4	15–30
PWP Clinical Lead	Expert	2–4	5–15
PWP Clinical Lead	Intermediate/ Expert	4–5	5–15
PWP Team Lead	Expert	4–5	15–30
Innovation & Service Improvement within Trust	Expert	5 or more	5–15

the vast majority (n = 11) having at least 1–2 or more years of experience in using SilverCloud as an iCBT service. Table 2 further demonstrates the case load of each supporter at the time of our design sessions, which most often included 15-30 (or more) SilverCloud patients, and this typically reflects ~50% of a PWP's overall case load.

4.3 Procedure and Data Analysis

All design sessions were conducted remotely via video conferencing software (Microsoft Teams). Informed consent was sought in writing prior to the study. The session began with the researcher (AT) *setting the scene* by describing some of the unique opportunities that are afforded through

digital health services like SilverCloud that can collect patient interaction and treatment data at scale, and through this, enable new possibilities for statistical analysis, including advanced computational methods such as AI and ML. We explained how AI methods can help recognize specific patterns of people's behaviors from large-scale data. This includes the identification of patterns in mental health symptom trajectories from clinical scores of PHQ-9 and GAD-7. Thus, looking at historic data of SilverCloud patients who previously completed iCBT treatment and regularly reported clinical measures, we can distinguish those patients who achieved a RI in their mental health at the end of treatment, from those who did not. An algorithm then learns the differences in mental health trajectories between those two groups (reliable improvers vs. no-reliable improvers) and can predict, early within treatment (i.e., by the third patient review), with fairly high accuracy (87%) whether, or not, a particular patient is likely to achieve a significant reduction in their depression and/or anxiety symptoms.

Presenting this as a concrete example of what AI could do, the researcher then explained the purpose of the design session as a three-part: (i) *inviting feedback on the proposed idea of predicting RI* to clarify use cases and the potential value of the suggested AI application (when, where and how, if at all, having access to RI prediction could be useful); (ii) *reviewing a set of design proposals* to better understand information presentation preferences and the integration of the data-insight within the existing supporter interface; and (iii) *discussing any concerns about the AI design concepts or their uses* in this specific mental healthcare context.

Each session lasted 1 hour and was audio-recorded. All recordings were fully anonymized and subjected to full transcription by an accredited transcription company. The transcripts were carefully checked by the lead researcher for the correctness and subjected to Thematic Analysis [12]. This involved an intensive familiarization with and coding of the data, and their iterative organization and development into high-level themes. Our analysis was guided by our three main research aims of: *clarifying use scenarios and the utility of the proposed AI*; *learning about concerns*; and *gaining feedback on the designs.*

All participants received a £30 gift voucher to a retail store of their choice to compensate them for their time spent in contributing to the research. The research study was carefully reviewed and monitored for compliance and privacy regulations; and approved by the NHS Health Research Authority (HRA, reference: 19/LO/1525). Each participant has been given a unique identification number to protect their anonymity, reported as P1–P21.

5 USER RESEARCH FINDINGS AND IMPLICATIONS FOR DESIGN

Next, we present the key findings that emerged from our analysis, which concludes with implications for design (Section 5.3.1) that were then taken forward into a first concrete UI proposal (Section 5.3.2).

5.1 Use Scenarios and Proposed Utility of Outcome Prediction to Supporters and Health Services

Having introduced the supporters to the prospect of a RI prediction, we asked how they would imagine making use of this outcome data in their day-to-day work. Explaining the potential value of the prediction, they described two main uses. They regarded *outcome prediction as a helpful indicator*: (i) *for adapting patient treatment for improved outcomes*; and (ii) *in performance monitoring*.

5.1.1 Adapting Patient Treatment for Improved Outcomes. In keeping with the clinical literature on FIT, having access to the prediction was perceived as a useful indicator for assessing patient mental health progress by providing an additional perspective—based on data. In particular, supporters regarded a positive prediction (RI) to serve as "confirmation" that a patient is "on track" and to add "confidence" that they can continue with their chosen treatment approach. In these

instances, the data was considered to "back-up what clinical scores may already suggest", providing reassurance to PWPs that the treatment is working. P13 for example explains:

> *(…) I think it would be a really good boost if you found out that, actually, this client should be on track and that they should be likely to reliably improve. I think that would be just quite a nice boost that, 'Oh, good. This is really going to be something that helps them'.*

Contrarily, supporters assessed a negative prediction (No RI) as a prompt to explore why the treatment may not be working for a patient, and how it could be adjusted to better meet their needs. To this end, PWPs described taking a number of different actions. They would engage in more conversation and "funneling" with the patient to better understand why they might not be on the right path, what they find (un)helpful and discuss any changes that might need to happen. Related to this, supporters would work harder to identify any treatment, engagement, or mental health barriers that the patient might encounter through a closer review of available data; as well as increase their level of support for patients who "need a bit more attention in the review" through activities such as: switching the patient from online to telephone reviews; breaking down therapy contents more; providing access to additional resources; increasing the relevance of treatment contents (via tailoring); adding more treatment sessions or increasing the frequency of review sessions. P18 explicates:

> *I think with an online review client in particular if I was able to tell, say if we had this data really early on and we could see that it wasn't going to be effective, then what that will allow is actually I could call that client, e-mail them, have that more of a verbal discussion to see what the issues are, how can we support them in a better way. Whereas previously I might have relied on the information that they're leaving me to highlight if it was or wasn't working (…).*

Furthermore, a negative prediction could also lead supporters to re-assess the fit of the chosen SilverCloud program for the patient or whether the therapy itself needed to be changed (i.e., to a face-to-face format) such that the patient would get the right kind of help sooner. For this, supporters described the benefits of using the prediction as data evidence for guiding decisions within CMS:

Firstly, a negative RI prediction for a patient can *support decisions if a patient needs to be stepped-up to different care "sooner"*. Regarding the prediction as additional, more "solid, objective data evidence", supporters assessed it as helpful for building the rationale and backing-up their clinical judgment to change treatment; especially if the prediction aligns with their intuition that the patient might not be benefitting from current therapy. This provides opportunities for improved care planning through an earlier re-distribution of care resources and reduced delays in having a patient attend for too long the wrong care pathway. P6 states:

> *Just hopefully the chance to change it. If you think that person's not going to get into like reliable change there, and identifying the correct treatment options, you know, the earlier on that you do that, the more likely treatment is to work, so if it was flagging that actually, yeah, when you've discussed it with your supervisor this maybe isn't the right option, it gets them into the right one quicker. Yeah, it just makes it a bit more personalized I think, as well.*

Secondly, the prediction can also *assist in decisions about how many (more) treatment sessions supporters should administer to a patient.* A positive RI prediction can confirm that the usual 4–6 review sessions might suffice, whereas a negative RI prediction as early as review session 3, would

enable an earlier change in intervention; as well as indicate that patients may need a longer treatment period if the change is unlikely to be expected within that time; offering useful insights to resource management. P15 expands:

> *I think at Step 2[5] with PWPs the most useful thing would be to not overuse sessions. So if you could predict early on that a patient at session three isn't going to experience RI it could change the intervention more quickly. So at the moment you might have a PWP who's working with a patient and they might do six sessions before they then move them to the next intervention, but if you can tell that somebody is not going to benefit or get to recovery from that intervention you might want to think about changing it sooner. How I see it would be useful would be in the step-up or changing intervention process.*

Despite multiple proposals for how especially a negative RI prediction can provide a useful indicator that a patient is not "on track", it does not provide specific, diagnostically-useful insights that could assist supporters' understanding of potential treatment problems, nor does the AI provide any therapy specific recommendations (i.e., for content tailoring). P1 describes this limitation and a desire for potentially more "actionable" data insights:

> *"I suppose the difficulty with a prediction is it is ultimately a figure that we're going to be receiving. And that doesn't actually give me detailed information about my practice. That's where I get a bit, because it is just a number, so I might get a number saying the prediction is, there's a 63% chance this person will improve and recover, great. But that doesn't tell me what I need to do. (…)."*

In summary, supporters described multiple ways in which they envisioned the proposed outcome prediction feedback may serve as a useful "flag" for reviewing their practices that would allow them to adapt treatment choices. This approach to personalization contrasts with other possible tactics whereby the AI could, i.e., help identify specific problems that a patient may encounter and provide to the PWPs specific suggestions for action. While potentially more insightful, we will discuss later our choice to pursue a simpler data insight in lieu of more complex inferences and conclusions that could be drawn by the AI to reduce risks that may otherwise arise in cases where the AI prediction is false; as well as to not interfere with supporters "sense of agency" in making their own professional assessments nor reduce opportunities for upskilling novices.

5.1.2 Early Indicator in Performance Monitoring. While supporters saw the most potential in the use of outcome feedback for treatment adaptation, some envisioned its use in *performance monitoring*, by providing insights into (i) the *effectiveness of individual PWPs*; (ii) *how well the health service is achieving its treatment targets or required resourcing*; and (iii) *anticipated success rates of a digital approach to help the uptake of online patients by supporters new to services like SilverCould.*

PWPs in team lead and supervisory roles described how the predictions could provide "really specific data" on how well their supporters are doing, potentially allowing them to better monitor and assist staff performance. In this instance, negative RI predictions are regarded as a "training tool" that provides supervisors with an opportunity to work more closely with a PWP to help move forward their thinking and decision-making in how best to support their patients; and to do so in

[5]The stepped care model used by IAPT for making clinical decisions: https://www.mhm.org.uk/pages/faqs/category/stepped-care.

a very sensitive and carefully scaffolded manner to ensure they empower (rather than criticize) the PWP. P6 explains the benefits of identifying guidance or training needs early:

> *Yeah I think it would actually end up having quite a big impact in that you'd just be making sure that your PWPs were doing the right stuff and had enough support, because I know that, for example, at the moment the only thing we can really go by is recovery rates but that's often at the end, or just waiting to see if people complain about your PWPs, which isn't nice! Whereas I think if you could see on SilverCloud that actually there were quite a few people who maybe weren't quite getting there it's just it would encourage me to then have that gentle conversation about 'do they feel like they know enough about SilverCloud', 'are they supported enough', and then could in theory make more clients get into recovery (…). This would give me like really specific data which would be really helpful.*

Furthermore, a few supporters considered the value of the predictions not just on an individual supporter basis but for providing a useful "helicopter perspective" to senior management and coordinators on how well the specific health service (i.e., NHS trust) is performing. Interpreting the prediction as a "forecast" on key performance indicators, they described its potential in clarifying if a service is due to hit its targets; or whether additional staff may be needed due to an identification of, for example, large numbers of complex patient cases.

Similarly, as a measure of service performance, the prediction is regarded as providing supporters, who are new to SilverCloud, with important insights into its effectiveness as a treatment. Especially, where predictions are predominately positive, PWPs described the prospect of it offering encouragement to clinicians who may be hesitant to get started with a digital treatment approach. P18 explains:

> *But I think this will be really valuable for people who, well clinicians who are new to the service and new to using this treatment because actually that will help build their belief I think quite early on that it does work, whereas I think I have that experience to know that it works and it works with a lot of people so I've got a lot of practice evidence that it works, whereas some people who've not worked in the service or not worked with SilverCloud may be a bit more hesitant of the digital work.*

To conclude, supporters described the potential benefits of the prediction as an early indicator in performance monitoring to enable more timely adaptations to their practices; better resource planning at a service level; and for communicating the prospective effectiveness of SilverCloud to supporters new to the iCBT service.

5.2 Potential Risks and Concerns about use of Outcome Prediction in this Mental Health Context

Next, we describe key concerns that were raised by supporters throughout our conversations of specific use cases and when reviewing the various design concepts, describing risks pertaining to (i) *PWPs feeling demoralized to take action and experiencing increased performance pressures*, especially in response to negative RI predictions; and (ii) the potentially *uncritical treatment of, and over-reliance on, the prediction outcome*, which may lead to *reduced PWP support* and *lesser engagement with, individualized, patient-centric care.* Discussing such concerns and how these could be addressed, we close with supporter reflections on (iii) *the need for balanced data* assessments that consider the AI prediction outputs in the context of other patient information; and (iv) the *implications of "false" AI predictions* for patients, supporters and care services.

5.2.1 Negative RI Predictions: Demoralizing Supporters to Take Action and Increasing Performance Pressures. While a positive RI prediction can boost supporter confidence in their treatment choices; there were mixed results in supporters' responses to negative RI predictions.

On the one hand, they describe negative RI predictions to serve as an important indicator for treatment adaptation and as a positive "challenge for them to improve the patient's situation", providing them with a "push" to explore more how to improve the patient's mental health. With this mindset, they ascribed negative RI predictions not necessarily to their own work practices but attributed it to limitations of the "online" medium; other "situational patient factors"; and possibilities of the "data not being 100% accurate". P4 explains:

> *(…) it's data. It's not going to be 100% accurate, because it's going to miss out little things. I wouldn't let it dishearten me, because I ultimately have that contact with the client and get to hear their feedback themselves that there may be small things that are stopping them from getting to reliably improve, like situational things. I guess I just wouldn't let it impact on me.*

On the other hand, the majority of supporters also raised concerns how access to such data insight might "negatively impact on assessments of their performance and competence as a therapist". Here, predictions that their patients are not going to improve as desired can contribute to feelings that they may not provide "the right level of care"; or are "not doing a good job". P9 states:

> *(…) you might take it as a personal, not attack, but a personal question mark over your competencies or why they're not going to improve (…)*

In instances, where negative RI predictions sustain for individual patients, supporters described how this could have a "demoralizing" effect and might lead them to give up. Expressing a reduced sense of agency and feelings that there is nothing they can do to change such outcome, P1:

> *(…) you've worked with them for three sessions and they're unlikely to make a RI, you just think, 'What's the point? Let's just wait another session and then I'll take it to CMS and deal with it', rather than actually being like, 'Oh, what's going on?'.*

Whether PWPs are more likely to perceive negative RI predictions as a "challenge to achieve improvement" or "thread to assessments of their professional competence" was described to depend on: (i) their *personality* or general *mindset*; (ii) their *workday* as determined by their caseload, associated time pressures and stress levels (i.e., high stress levels can increase sensitivity to take outcomes personally rather than to explore external factors); and (iii) the *frequency of exposure to, and distribution of, negative predictions across their patient cohort.* For instance, P1 explains how frequent exposure to negative predictions, especially in situations where they are very invested in patients, can impact their motivation and self-confidence:

> *I think that's where I feel this is tricky. This is where I feel a bit conflicted about how often we would see the stat, because I don't think it would… I don't think necessarily it would impact my level of how motivated I would be to support the patient, but more I would find it demoralizing. If I felt I was working really hard with that patient and I've reflected on it and I'm trying to do this, that and the other, and it keeps going back, I'm like, 'Is this me?' So, I guess the danger is you start kicking yourself constantly about it, after you've done all the exploration. So, I guess you try and pull out all the stops, yeah. But then there are kinds of gets that line where you're like, 'Okay, how constructive is this?' it's finding that balance, I think.*

This suggests the need for careful design choices as to where, when, and how often supporters should interface with the prediction to reduce risks to supporter agency, confidence, and motivation. This may include decisions to only show the prediction to PWPs at specific (clinically relevant) times in the treatment journey to aid perceptions of it serving as a "flag" for reflection on patient outcomes.

Furthermore, how the prediction is positioned and contextualized within the supporter interface can invite patient outcome comparisons and additionally contribute to performance pressures. For example, the Client Dashboard (Table 1, concept 5) shows the prediction outcome for the entire patient cohort of a supporter. While this can help identify those at greatest need for more support, it depends on the distribution of negative RI outcomes whether a supporter feels motivated or demoralized to affect positive change. P18 explains how a predominance of no-RI predictions may feel threatening to self-perceptions of their professional capabilities:

> *(…) [If] I was to see that actually there was quite a low likelihood of all those people in the top 10 that I've got listed as soon as I logged onto SilverCloud every day, I think it would demotivate me but also, I think if I knew that the algorithm suggested it wasn't going to work it might affect the way that I approach that client if I could see that all my clients weren't going to work. I think it would definitely filter into some negative thoughts about my work practice abilities!*

Similarly, supporters raised concerns about the Population Comparison (Table 1, concept 4), which was seen as another way to potentially monitor and compare supporter performance. For example, in instances where a patient is predicted to not improve and is less likely to improve than the general population statistics suggest, this could be seen as an indicator of the supporter underperforming. This can add to performance pressures within a work context that already "strongly emphasizes statistics and numbers to justify pay' as well as invite 'unhealthy competitive thought" whereby supporters might start to compare themselves to one another, which could detract from the important focus on the patient. P20 describes:

> *I think with the population comparison it comes back to that kind of concern around that bit of competitiveness (…) you might start thinking, 'Well, I'm not doing as well as this person', and, 'They've got scores of lower than majority' and that just might have negative implications for me as a practitioner because I'm thinking, 'Oh, I'm not doing that well'. I don't want to feel like I'm doing a crappy job. (…)*

All this suggests for design to focus supporter attention on an individual patient's care for improving that person's mental health outcomes, rather than to encourage cross-patient or cross-supporter comparisons.

5.2.2 Uncritical Treatment/Over-reliance on Prediction Outcome. High caseloads, fast-paced work, and emotional stresses can mean supporters are not able to pause and reflect about a prediction in a "balanced manner" and could become overly reliant on it. This could manifest in PWPs uncritically accepting and over-trusting the data, which some may regard as a useful "shortcut" to making "quicker" decisions; potentially disregarding their own reflections, and what they know about the patient. P3 expands:

> *Well, our job is really fast paced. Most of the time it is hard to allocate a time to actually reflect of where this client is now, where we're heading, what do I do with it? And I think the downside of this is that therapists will disregard that reflection and just rely on the machine. 'I'm going to just click on this and see what the machine tells me', and, like I said, just disregard that side of reflection whether is this right based*

> *on what I know so far myself and having that relationship with the client? Do I feel that this fits in the situation? I guess that will depend from therapist to therapist, but then, like I said, our culture being fast paced, we can easily be drawn into shortcuts, give me the answer.*

This tendency for over-reliance is moderated by assumptions that high confidence predictions provide "good data" that is "based on tens of thousands of previous cases", which many PWPs described as potentially a "more reliable", "more objective, evidence-based" indicator than their own (human) judgment. P16 explains:

> *(…) I know my intuition is not perfect and I know the data is not perfect, so it just opens up those conversations, doesn't it, which it needs to, and I'd be inclined to go with the data most of the time anyway, unless I've got a good reason not to. I'd probably trust the prediction accuracy as long as it was good predictability, I'd trust that more than my own intuition unless I had a really good reason not to.*

Individual assessments of patient outcomes can also be more difficult to develop: (i) early on within treatment; (ii) for more complex cases; and (iii) where PWPs are less experienced in their work (i.e., trainees, novice PWPs, and other non-trained practitioners using SilverCloud), which increases chances that PWPs uncritically accept a prediction outcome "as is". Interlinked with a potential over-reliance on the predictions are concerns about *reduced support* and possible *neglect of the more individualized side of care*, which we will expand on next.

In a few cases, supporters remarked how seeing a positive trend may suggest that their help could be less intense and take more the shape of "check-ins" with the person; encouraging them to continue with what they are doing; and shifting the focus to the online program as the main source of support as opposed to the PWP. In other words, some PWPs may—consciously or unconsciously—reduce their level of support for patients predicted to improve, assuming their recovery and use of SilverCloud in a self-sufficient way. While some supporters describe the benefits of being able to expand their "scope to help those patients who are unlikely to get to RI", others expressed worries that over-trust in the prediction could mean some PWPs make rushed decisions to change or not change treatments without considering the individual case as much. P9 expresses:

> *I'm just scared that there might be negligence in the sense that therapists may become too reliant on it, that they may … 'Oh well, they're going to recover in four weeks' time anyway, so let's just quickly push them through the treatment' and not really look at longevity and making sure that there is going to be … (…) I don't need to do this bit or that bit or we can skip this corner here or if they need it, it's always going to be there available to them in SilverCloud.*

In such instances, the prediction is seen to potentially "disrupt" the usual efforts of PWPs to engage with patients in dialogue to identify how best to support them, or to gather additional information about the patient that could enable a more balanced assessment. P3 states:

> *This is what I'm saying as a disadvantage. You then lose the interpersonal side of things of have I checked if this person can actually then do it on their own without offering two more sessions? Because it can be just reliant on let me check in with the machines and see what he says about this.*

All this emphasizes the need to ensure—through appropriate PWP training, supervision, and continued patient monitoring—that interpretations of the prediction should not mean a reduction

in care that could risk having adverse effects on patient recovery, nor to forget about the all-important individualized side of care.

5.2.3 Need for Balanced Data Assessments: Imperfect Prediction Outputs are Only "One piece of Data". In keeping with the above, the majority of the supporters themselves were very mindful of and expressed the need for, the prediction to not take away from the "human-centred care" that is tailored to the individual. Cautioning that patient scores alone—which the predictions are based on—may not accurately depict a person's mental health, many supporters reiterated the importance to not treat the "patient as a number" and put clinical decision-making as "black and white"; advocating instead for supporters to use their professional skillsets and talk to the person to understand their struggles.

Furthermore, the vast majority of PWPs generally acknowledged that they would not fully rely on, or 100% trust the prediction at all times; describing their intent to treat the prediction as "one bit of information" that can serve as a "helpful indicator or guide", but should not be treated in isolation. Instead, PWPs stated that they would evaluate the prediction in the context of other (i.e., situational) factors that influence, change, or otherwise explain patient outcomes. Considering the often unique circumstances of each patient, they explained how the main rationale for their any decision-making should still be based on what they are picking up from their patient, and their own clinical judgment. P6 describes:

> *(…) So I think it gives a bit of backup to our rationale I think as long as we're using it to back it up and not just as its own thing, then I think that would be helpful. (…) especially as we can never 100% predict what someone's going to do.*

Their feedback suggests that incorporations of the prediction within their work practices should nurture a more symbiotic relationship between the data and clinician assessments. To supporters, the prediction presents "a suggestion based on a pattern"; and they regard the recognition of that pattern in itself to have value as it enables them to then pick up and pinpoint extra aspects in their care with the patient that the machine may have missed in its data. P4 articulates this as follows:

> *(…) it might be missing certain things. It's not going to pick up everything, and we know it's not going to be 100% accurate. It's going to be as accurate as it can be. But I guess it's just useful, isn't it? Because we as therapists are able to point out those things that are maybe not in the patterns, or pick up on those things that are not in the patterns as we talk to our clients. That would be my only concern is maybe not picking up everything and including everything in that pattern, so I probably wouldn't rely on it 100%.*

To achieve a more balanced, symbiotic relationship in the way that supporters should come to work with the predictions to mitigate identified risks, requires careful PWP training, explanation, and interface design to effectively communicate what may constitute an "appropriate" data use: what is the prediction for; what are its limitations; and how can it be (un)helpful to supporter practices?

5.2.4 Implications of "False" Predictions. While PWPs generally assessed 87% as a good prediction accuracy, they were mindful that the 'machine could be false', cautioning them to not treat the prediction "as a fact". When asked during the interview how supporters would feel if they discovered predictions were false, they described this as 'irritating' and to cause them to become less trusting and more dismissive of the data, especially if it was false often. Discussing possible implications for patients, they considered false predictions however to only have a "small impact". They ascribe this low impact, firstly, to the use of the prediction tool in conjunction with other

available patient information (as detailed above). Secondly, PWPs highlighted the importance of continued monitoring of treatment outcomes following any adaptations made. Reflecting about the consequences of a false prediction, P13:

> *(…) We're still going to look into ways to still support that client, even if we don't think that it's going to be of any benefit. We're still going to look into barriers and how to still make the best of it and get the most from that programme, even if it's not going to be the sole thing that [this] person needs. So, I think if it's incorrect, I don't suppose it would make too much of a difference, because like we said, we're monitoring them anyway. So even if it says, 'Yes, they're going to be fine'. Like that one says, '98%, it's going to improve, you've got nothing to worry about', we would still be monitoring that anyway. So I don't think it needs to be 100%.*

However, false predictions were also described to be disruptive to a patient's treatment journey if they caused unnecessary or unhelpful adaptations that could lead to poorer patient experiences or meant that the patient misses out on support needed or be delayed in their recovery. About the importance of correct predictions, P20:

> *I think it would have to be very important because we're kind of getting to the point where we're drawing a line between helping somebody for what they need versus just going back and forth between different things because we think that's okay and that's the right thing to do. Which I'm happy to do if they're going to benefit from it, but if that information wasn't that accurate I'd feel really bad for the patient because I'm just jimmy jamming back and forth for no reason really, or not a good enough reason.*

Interlinked are concerns how prediction errors may falsely represent the performance and anticipated recovery rates of the health service, which can impact future funding. Prediction errors may also lead supporters to administer more treatment sessions than needed, impacting their caseload and care capacity as well as potentially adding to supporter fatigue and burnout.

Despite those concerns, supporters generally assessed the benefits of predominantly correct predictions to potentially outweigh risks of false predictions due to the prospect of overall improved decision-making and achieving better patient recovery rates than what they would achieve through current practices. P16 explains:

> *Well it can only be correct X% of the time, right? So it don't bother me if it gets it wrong, it's telling me it's going to get it wrong. It's just the best that it can do based on loads of data, which I don't have otherwise. (…) machines make false predictions, it's just saying nine in ten people, so one in ten times, it's going to give you a wrong prediction. (…) for that individual client, it's a pain but that's IAPT, you have to make these decisions anyway. At least you're getting it right 90% of the time now instead of, you know, 60%. (…) it's about improving my decision making, isn't it? Can this improve my decision making by 10%? If it gives me a false prediction, but overall, I know, my recovery rate increases by 10% and clients are getting better, then I'm happy because patients are always not going to get well.*

Thus, while overall risks to patients and their care might be low, supporters cautioned how high rates of false information through prediction errors could negatively impact patient experiences; health service outcomes; and supporter care capacity. Simultaneously, we find that their feedback reflects a very balanced assessment of the risks overall through expressions of awareness that AI predictions will not always be correct, and suggestions that—on a large-scale service

delivery level—the combined benefits of sufficiently accurate predictions are likely to outweigh the risks. Remaining mindful that limitations introduced by prediction errors and model uncertainty are difficult to overcome fully, it is nonetheless important: (i) to carefully explain to, and remind supporters—both in PWP training and within the user interface—that prediction outputs are probabilistic in nature, and thus, can be fallible; and (ii) to encourage them to contextualize and balance the prediction outputs with their own professional assessment as well as other information about their patient.

5.3 Design Feedback and Implications

A summary of how supporters assessed each design concept is included as Appendix C. This section extracts the key design learnings from the feedback to enable the integration of the AI models into a production interface. We chose to incorporate this level of detail since the specific design choices made for AI applications are rarely reported in existing literature, yet, this design and implementation stage is an essential step forward in any efforts to move towards real-world deployment.

5.3.1 Identified Requirements and Sensitivities for Design. In reviewing and comparing the feasibility of the design concepts, we identified four key sensitivities for design:

Create a context-appropriate visual that is easy-to-interpret at a glance: Due to busy schedules, the supporters unambiguously expressed the need for the prediction representation to be very easy-to-comprehend and quick-to-interpret. Here, many valued the symbolism of the Wheel Indicator whose graded chart with its green and red areas enables a "near instant understanding of a lot of information". Yet, one supporter perceived the "red NO" label as a too "stark, attacking" visual, describing it alike a warning signal in cars that "screams failure" or "danger". There was also potential for confusion if the Wheel was read as showing the prospective spectrum from deterioration (far left) to improvement (far right) rather than as a binary (yes/no) indicator for RI. As a more "neutral", non-threatening, and simple design, the Visual Cue + Text (concept 3) felt more appropriate and was received best. It was assessed as a "clear, succinct and easy-to-access" presentation that solely focused on the RI prediction metric. The other concepts of Multiple Outcomes Table, Population Comparison, and Client Dashboard, all entailed more information and complexity that, whilst generally providing additional insights, were regarded as "too overwhelming" and to take up "too much time to read" and unpack. Given the constraints on supporter time, all this suggests the need for a fairly neutral, simple design with low information complexity. This necessitates a difficult tradeoff between providing as little information as possible whilst ensuring that the AI prediction becomes correctly understood for what it is, and appropriately interpreted for informing treatment. To manage information complexity, PWPs also recommended to communicate only "one concept" upfront, with the option to include other outcome definitions or additional context-information "on demand".

Cater for diverse information representation preferences to help map prediction (category) to action: Overall, preferences for representation modalities were mixed and varied considerably across the supporters. In keeping with the above-described importance of time efficiency, most PWPs preferred designs with simple, clear visuals (concepts 1 + 3) that enable them to "easily see rather than having to digest the information of numbers" or text. Least liked were communications of the prediction output through text alone, as in the Multiple Outcomes Table. The text labels were assessed as "too vague"; making it difficult to understand the nuances or categories that distinguish, i.e., "very likely" from "likely" labels. This often led to expressions of a clear preference for, and proposals to add, numerical values. Especially for the Wheel indicator, supporters explicitly expressed their liking of the confidence percentages above text labels and simpler yes-no

indications. Some, however, also raised concerns about how these numbers are to be interpreted, describing the potential to get hung up on the detail of these numbers, and to not have time to reflect about them under time pressure. P9 explains:

> *(…) I feel like I'd get too hung up on, say for example, if it's 95 versus 98, I'm like well what's that 3%? Whereas if you have a word, it's objective. If you can see where I'm coming from. It's less subjective, like well okay, this one's only 58, that's fine but it could be 60 but it … so you … it's hard …(…) if you've got a word that's objective, you interpret it, you understand it. It's a universal indicator, clearly. I appreciate numbers are universal as well, but unless we've got a clear scale, you don't really know what you're working with.*

Whether supporters found text, visual or numerical information easier to process depending on their individual preferences and ability to map the prediction result to a "category" or "clear scale" that allows them to make sense of the data in their mind such that they can link it with subsequent action. While for some, text labels can feel less ambiguous than numbers, it is reverse for others; suggesting a hybrid design that includes multiple formats to account for diversity in preferences and aid meaningful matches of prediction (category) to action.

Facilitate PWP work practices in providing patient-focused, individualized care (not work pressures): Preferences for the various design concepts were, of course, also moderated by the anticipated use scenario and how the prediction data could potentially be actioned. Most values were ascribed to the use of RI prediction in assisting the individualized care practices of PWPs by enabling targeted adaptations of the patient's treatment for improved outcomes (Section 5.1.1). To this end, design concepts that showed the prediction in the context of an individual patient's profile were preferred over, i.e., the Client Dashboard (concept 5). While the dashboard enables the spotting of trends in supporter performance across patients, which some supervisors regarded as useful insight; others raised concerns about it potentially reflecting badly on their work practices, especially if many patients are predicted as unlikely to improve. What might complicate these reviews is that the predictions shown in the table reflect the outcomes of patients at different stages in their treatment journey, which might be less appropriate to compare across. Similarly, for assessing an individual patient, the Population Comparison (concept 4) was evaluated as "least helpful". While some supporters saw the benefits of being able to give oneself a "pat on the back" if the visual showed for their patient a higher chance of RI than for "similar patients"; this was a concern for others, who saw additional statistics that showed comparisons as adding to supporter stress and competition, and who questioned the relevance of such comparison for aiding individualized, patient-centered care or discussions of an individual patients' circumstances during CMS meetings. P1 states:

> *(…) And as a practitioner, personally, I don't think that's helpful for how I would treat my patients because it's hard enough personalizing treatment already, let alone being like, 'Oh, well, compared to others who are in the same situation, they should be...' And people shouldn't be fitting in with everybody else. That's the whole point of mental health. It's not about should, it's about 'where are you'. So, I wouldn't want to see that comparison.*

All this re-emphasizes the importance of supporting "individualized care" practices and avoiding threads to supporter motivation or adding to performance pressures through patient comparisons.

Help ground prediction outcome in real-world data: Despite the above critique of the Population Comparison concept, it is noteworthy that some supporters assessed the additional information it provides to offer a useful context for interpretations of the prediction and treatment choices.

For example, a less than average chance of RI may influence how closely a supporter monitors or supports a patient. Here, some of the PWPs suggested that the comparison information "gives more rationale to why the machine has decided about the chances of RI" and "more explanation why the data should be considered". In other words, some of the supporters described how having this additional context information can help ground the prediction within real-world data, and how this can help their understanding of, and trust in, the AI-generated data. P6 explains:

> *I think it's just a bit more interesting than just having a prediction, I think that's really good but because it's drawing on real life examples, it's taking you back to people who have actually used the program rather than it just being a number. I think it makes it a bit more solid.*

Thus, while patient/ group comparisons may be less appropriate for this specific mental health context, and more complex information likely ignored due to work stresses, providing nonetheless access to information (on demand) that details the source data may aid to the credibility and potential acceptance of the prediction within clinical care. For this, it is paramount that the design aim remains to encourage understanding of the prediction rather than to persuade supporters to uncritically accept the data output.

5.3.2 Design Rationale: Choices Made in a First UI Integration for RI Outcome Prediction. Responding to these findings, the UX team at SilverCloud designed how the RI improvement prediction would interface within the Supporter Intervention Management site, as is shown in Figure 4.

Patient-focused Indicator: Aiming to serve as a *useful indicator to inform and potentially adapt the treatment of an individual patient*, the prediction is placed prominently within the "patient review status" section at the top of the patient profile page. This enables supporters to quickly assess how actively engaged the patient is with the iCBT program; whether they are due a review; and what their prognosis is for achieving RI. This placement allows for the prediction to be seen and assessed in the context of all other patient information, including the trajectories of their depression and anxiety scores over time, which are the inputs to that prediction.

Succinct, Clear and Diverse: A key driver in the design is ensuring that the RI prediction can be understood quickly, at a glance, by a non-technical audience with varying levels of expertise as PWPs and varying experience in using SilverCloud. Building on the Visual Cue + Text, the design charts five boxes which map to a five-point Likert scale that is articulated through short text labels that categorize the prediction as a range from "very unlikely" to "very likely" for a patient to achieve RI. Color gradients of grey-to-green help create a neutral, non-threatening visual. To not overcrowd the design and unnecessarily draw too much attention to the detail of numbers, percentages are included in a drop-down menu to account for diverse information review preferences.

Text Clarity and Explaining Non-Predictions: To ensure that the words chosen for labels and contextual explanations are precise, help an accurate understanding of the prediction, and reduce risks of the information being misinterpreted, requires multiple iterations and considerations, including whether to provide explanations, i.e., for "empty state" conditions for which we discussed that it would be better to not show a prediction. This pertains to (i) patients, who have not yet completed three clinical questionnaires required to make an accurate enough prediction; (ii) those whose starting PHQ or GAD are below the thresholds for RI (PHQ-9 score <=6, GAD-7 score <=4) and thus, achieving RI is numerically impossible; and (iii) those whose score(s) are below caseness.[6] For patients below caseness, it is numerically very difficult (albeit possible) to

[6]Caseness is the clinical cut-off score (PHQ-9 of >=10, GAD-7 of >=8) that indicates whether a person's symptoms are sufficiently severe to be considered a clinical problem. Patients with scores below that threshold are considered as non-caseness or recovered [21].

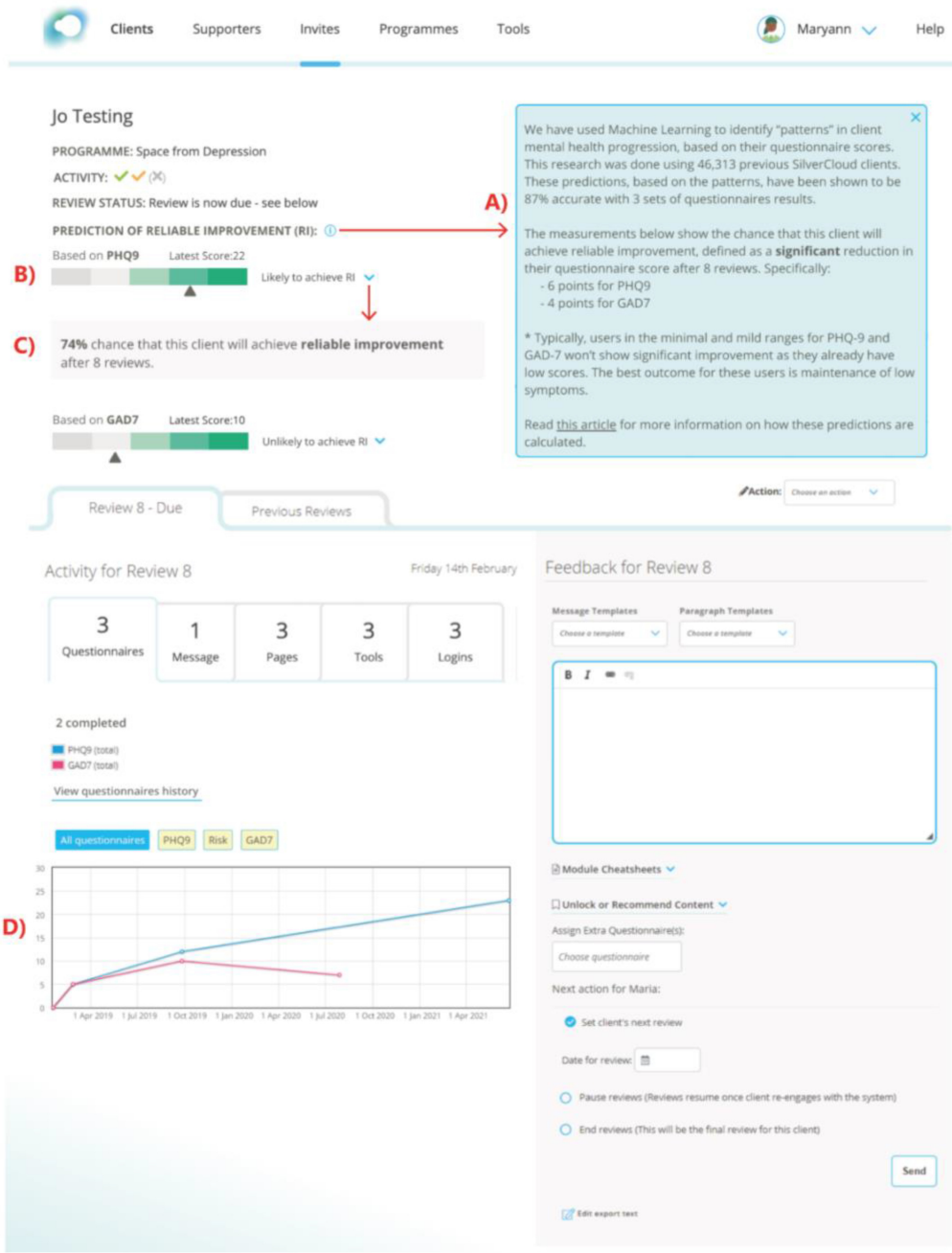


Fig. 4. Integration of the RI prediction within the patient status of the supporter Intervention Management Site, showing: (A) a general explanation of the prediction; (B) visual charts with a text label to convey the prediction results for PHQ-9 and GAD-7; (C) drop-down menus for numerical percentages of the prediction; and (D) other contextual information about the patient that is considered in their review, including their clinical score trajectories over time.

achieve RI, which makes negative RI predictions highly probable. Yet, those predictions are less likely to indicate poor patient progress or need for more support, as these individuals are already in the desired score bounds of "recovery".

Explanation vs. Openness for Interpretation and Appropriation: We also discussed including additional explanations for patients in the minimal-to-mild score ranges to explain that these "typically won't show significant improvement" and to suggest that "the best outcome for these users is the maintenance of low symptoms". While there can be the desire to *provide extra information for cases that can be more difficult to interpret*, we decided to include those only within a general but not instance-specific explanation to avoid *over-interpreting individual cases for* the supporters. These are deliberate considerations in order to: (i) avoid for PWPs to ignore the prediction unless it was for more serious cases; (ii) nor to restrict how PWP may make use of the prediction within their clinical practice. Partly, our research seeks to better understand exactly their clinical decision processes as well as any unanticipated use scenarios for the prediction, especially for unusual patient cases. Given the often unique, specific circumstances of each mental health patient, to generalize or narrow interpretations of the prediction, or to suggest what (concrete) actions to take in response, may risk reducing the applicability and potential usefulness of the data insight for a wider variety of patients. It could also feel disempowering to supporters, and impact their *sense of agency*, if they felt the prediction (and proposed actions) would try to replace their professional assessments, or otherwise limit their choices.

6 DISCUSSION: HUMAN-CENTRED AI IN MENTAL HEALTHCARE

In this article, we described our collaborative, human-centric approach to developing two AI models that can predict early-on, if a patient undergoing iCBT for depression and anxiety is likely to achieve a RI in their mental health symptoms by the end of treatment. We detailed how user research with iCBT supporters provided key insights into their work and information needs and how this, coupled with insights from the clinical literature and data availability constraints, enabled us to identify useful AI application scenarios and development targets for this mental healthcare context. To review our choices in pursuing outcome prediction and clarify the potential utility of the proposed models for clinical practice, we reported the findings of design sessions with iCBT supporters that investigated the integration of the achieved AI predictions within existing workflows. This further surfaced important concerns and risks associated with the use of outcome prediction in this context as well as a set of design sensitivities and requirements for developing appropriate representations of the AI outputs that resulted in a first UI realization within the SilverCloud product. Next, we will expand on some of these learnings and share our reflections on what constitutes a human-centered approach to AI design in this mental healthcare context.

6.1 Empowering not Replacing Clinicians with AI: Towards Human-AI Partnerships in Healthcare

There are many ambitious visions on how AI may drive forward health diagnostics, clinical decision-making, or treatment delivery, including—ultimately—the development of standalone AI systems such as the autonomous delivery of psychotherapy interventions. In such (future) scenarios, AI systems are often positioned as either capable of emulating humans (e.g., conducting health assessment, acting as therapist) or superior to humans, potentially outperforming them through improved data insights or productivity [101]. However, as discussed in recent literature [55, 70], it is unlikely for technology to achieve enough technical sophistication to replace human clinicians anytime soon. Thus, we believe that a more realistic, nearer-term, and perhaps more desirable strategy for developing AI applications is to orient design efforts towards the configuration of partnerships in how clinicians and AI insights might come together in healthcare delivery (see also [15]). Referring to the term "augmented intelligence", Johnson et al. [55] suggest that

while current AI does not replace humans, clinicians who use AI will replace those who do not. Miner et al. [70] further formulated four approaches to care provision: (i) human only; (ii) human delivered, AI informed; (iii) AI delivered, human supervised; and (iv) AI only—all of which have different implications for scaling-up care or ensuring quality of care. Thus, as a first, tentative step forward in introducing AI within an actual mental health service, we chose for our work to focus on the sensible integration of AI insights within human-supported care practices. Suggesting that those data insights can serve as a *useful resource for humans* [101], we discuss next our specific design goals: (i) for enabling iCBT supporters to build on (or extend) their professional expertise and protect their sense of agency; and (ii) to not unnecessarily interfere with the all-important "therapeutic alliance" between clinicians and patients.

6.1.1 Positioning AI-derived Data Insights as Inputs to Human Sense- and Decision-making Processes. Our user research identified two main ways in which predictions of RI outcomes could assist the work practices of iCBT supporters. They could serve as: (i) a "validator" to help confirm supporter decisions in cases where positive predictions align with their own clinical assessments, potentially boosting supporter confidence; and as (ii) a "flag" for negative prediction cases or where predictions were incongruent with supporter assessments, inviting pause to reflect and re-evaluate the patients' current situation that can prompt for adaptations to existing practices. As such, our AI output does not provide any more specific (i.e., diagnostic) information that could assist supporters understanding, e.g., of the patients' mental health state or potential treatment blockers, nor does it provide any concrete recommendations for what actions to take or propose to a particular patient. While more advanced AI applications are technically possible and could offer valuable additional insight, there are a number of reasons why we pursued a more general, less directive approach for generating AI insights:

Tradingoff the Risks and Benefits of Designing Complex System Inferences vs. Simpler Data Insights: Firstly, whilst the delivery of more complex data insights is an exciting prospect, it can be more challenging to achieve sufficiently robust and reliable data models. This is particularly pronounced in mental health due to general difficulties to establish what would constitute an optimal (aka ground truth) approach to treatment for a specific patient, even amongst health professionals [33]. In other words, while more ambitious algorithmic modeling efforts may propose greater gains, these can also come at an increased risk for cases where patients are falsely predicted for [101]. We believe that this presents a key challenge, especially for the design of *personalized* interventions that seek to increase the relevance and outcome of treatment for a specific individual. Yet, in cases where more specific, tailored recommendations may fail to deliver on their promise and mismatch the needs of patients or care providers, this can have opposite effects on patient engagement and health, and diminish AI utility and trust. Being mindful that AI systems are rarely, if ever, 100% accurate, we were very deliberate in our choices to explicitly position AI-outputs as part of human (expert) assessment and decision-making processes as a mechanism for managing those risks. In doing so, this leaves the human, rather than the machine, accountable for interpreting each patient's unique circumstances and, in response, determining appropriate actions forward. It also broadens the scope for other, potentially unanticipated use scenarios of RI prediction, and ensures its application to a wide range of patients. Thus, tradingoff risks and benefits, we consider this "AI informed" approach to human-supported care delivery [70] as a more ethical and responsible path towards early introductions of AI insights into mental healthcare contexts.

Designing for Human Expertise and Agency in AI-Informed Work Practices: Secondly, our research investigates how we can empower clinical supporters with AI. Thus, our aim is not to reduce the need for supporter input and analytical effort (in favor of the technology), but to explore how AI-insights could help maximize the impact of their "human" involvement in patient reviews. For this it is paramount that the supporters do not perceive the provision of AI-insights

as competing with, or replacing them in, their professional expertise as this could unnecessarily undermine them in their role; as well as reduce their willingness to support the development and adoption of AI approaches in their work [100]. Thus, by creating AI outputs that serve merely as a useful "flag" to inform clinical care, supporters remain "in charge" of examining more closely the circumstances and potential reasons for a particular prediction outcome and determining directions forward. The hope is that this can help preserve a *sense of agency* and *purpose in their role*, which is important for supporter motivation and job satisfaction. Other research exploring decision-support [52] goes one step further and argues for AI systems to explicitly suggest appropriate next steps within the technology design to help clinicians make the connection between AI output and their healthcare practices. Either way, for HCI research this suggests opportunities for interface design to—implicitly or explicitly—aid supporters in the identification of the right subsequent actions, which in our case may involve explicit design decisions to assist supporters in their search for explanatory information (i.e., by encouraging them to look for certain mental health blockers).

Understanding the Impact of Design Choices on Work Practices and Workflow Integration Challenges: Our study findings also identified how specific design choices such as the frequency of (especially negative) prediction outcomes or their positioning and contrasting with other information (especially comparisons across patients) could add to work pressures, cause demotivation, and a reduced sense of agency in supporters that their actions can indeed affect positive change. All this warrants careful considerations in future design and research to study the actual impact of achieved AI predictions: (i) how AI applications can help care providers make more-informed, confident treatment choices for improved patient outcomes; as well as (ii) how integrations of AI outputs within everyday healthcare come to shape clinicians understanding of their own role; and how their design can help to minimize disruptions to their clinical expertise and work culture (cf. [95]). All this can help advance learnings how AI technology may best assist health care providers in their practices.

6.1.2 Protecting the "Human-ness" in Human-supported, Digital Healthcare Delivery. While, as described above, there can be many different visions for how AI technology could come to transform (mental) healthcare, we have chosen to focus our efforts on identifying strategies forward for empowering (rather than replacing) clinicians with AI. Especially in the context of psychotherapy, we further acknowledge the importance for technology to not unnecessarily disrupt the interpersonal relationship between patient and care providers; seeking to protect the all-important "human touch", "genuine sense of care" and "empathic understanding" that often characterizes these relations, and are crucial for treatment success [92, 115]. However, as indicated in our initial user study (see [99] and Appendix A for main findings), trying to foster a connection between supporters and patients within a remote, self-administered therapy format that involves the asynchronous sending of online messages can already put into question the authenticity of supporters' identity as "real" humans. To counteract this, our findings describe supporters active work in carefully crafting their feedback messages to patients such that they convey a "sense of care" by including personable expressions; person-specific guidance; communicating that they heard the person's concerns; and ensuring that they respond to these concerns in an "empathic way" by building on their own life and professional experiences.

Integrating AI Insights Sensibly within Interpersonal Dynamics and for Supporting Human Relations: Given the importance of developing a *genuine* bond between supporter and patient within a computer-mediated setting, we were therefore deliberate in our choices for the AI and possible optimizations to supporter work to not go down routes towards standardizing or otherwise automating existing processes. Aiming to protect the "human-ness" of supporter communications we would favor, for example, the personal look and handcrafted feel of their personalized

messages that bring forward individual communication styles, over more templated, machine-led communication approaches that may increase efficiency in message production, but at the costs of inviting perceptions of a "robotic, auto-responder system". We believe that if we move beyond common development goals of "improving productivity" and considered more closely what may constitute a desirable use and integration of AI insight within healthcare from a patient and care provider point-of-view, this can open-up many important additional routes for AI application. Using goals of "protecting or nurturing supporter-patient relationships" as an example, future work may explore uses of AI to: (i) help increase patient awareness of the supporter's role and investment in their therapeutic success to foster their bond and associated benefits; and (ii) assist supporters to more closely connect with their patients. In this regard, our feedback-informed approach (RI prediction) in itself is sought to give supporters an additional view-point on their patients to enable them to be more responsive to those most in need of additional care, which can aid their therapeutic relationship. Future work may also focus more explicitly on the relational needs of the supporters by generating, i.e., data insights that foreground: how their actions came to matter to patients (i.e., highlighting support successes, or what types of actions are most helpful to their patients); what communication styles their patients may respond to best (see work by [24] as an example); or otherwise expand ways in which supporters specific skills and expertise can become leveraged more. Such efforts can aid a feeling of "congruence" on the part of the supporter for investing in the patient's treatment success, keeping them engaged and motivated in the process, which is often rooted in an underlying desire to "be helping others".

AI Acceptance vs. Perceptions of AI Dehumanizating Healthcare: Such considerations of where AI technology might come into interpersonal dynamics and caring relationships with a view to both sustain or extend human relations and avoid undermining health professionals in their roles and expertise may further play a key role in *improving acceptance of AI applications within such care contexts*. Especially in healthcare, there are increasing concerns about the role that AI might play in "dehumanizing medicine" [95]. Above and beyond already existing trends within health services to "continuously monitor" outcomes and focus on success "metrics", there are tendencies within AI work to treat individuals as "data points" in algorithmic modeling [20] by transforming a person's individual (mental) health experience into compressed mathematical representations that allow for the identification of large-scale patterns [100]. This is a tension that we also saw in our user research findings that highlighted concerns about the introduction of a binary prediction to lead to simplified interpretations whereby patients become treated as a "number" and prediction outcomes simply read as "black and white". Thus, in dealing with imperfect AI technology, as we will discuss next, it is paramount that we ensure in design and training that healthcare providers can maintain a more holistic view of their patients and focus on individualized care.

6.2 Dealing with "Imperfect" Technology in a Time-constrained Context: Implications for Trust in AI

In this article, we reported key concerns raised by iCBT supporters about the integration of AI insights into existing practices. This included the importance to avoid demoralizing supporters to take action and increasing performance pressures (i.e., by avoiding cross-patient or cross-supporter outcome comparisons); as well as multiple considerations pertaining to the interpretation and use of the prediction outcomes. Specifically, our work brought forward well-known risks and implications related to *prediction errors,* especially in cases where (more novice) supporters may *uncritically treat and overly rely* on the AI predictions; and where such reliance may cause *undesired changes to the intensity and nature of patient care.*

Moving Beyond Model Explanations: To better manage such risks, which are rooted in clinicians over-trusting the data, prior research in the field of **explainable AI** (**XAI**) suggests that

providing interpretable explanations of the workings of the model can help cultivate transparency and assessments of the accuracy of offered predictions that enable the development of a more appropriate understanding and level of trust in AI outputs [52, 105]. Yet, in time-constrained healthcare contexts, such as our iCBT setting, clinicians expressed their inability to engage with additional information. Instead, they emphasized the importance for the predictions to be understandable "at a glance" and that extra information—especially about the origin (or validation) of the model and how outputs are calculated—should only be available "on demand" to not distract from those insights most critical to their review and patient care (cf. [111]). This echoes other recent findings on CDS systems [52, 95] that describe how clinicians lack the extra time and mental capacity required to engage with such explanations, and that often assume substantial technical expertise and clinician interest in interrogating AI outputs. Thus, rather than a deeper understanding of how the AI insight is generated, clinicians favor an understanding of how they can make effective use of that information within their practice. Furthermore, Hirsch et al. [49] found that the willingness of mental health professionals to trust the AI output was bound up with the perceived "legibility" of the AI results (the extent to which the AI output made sense to the person) rather than the extent to which the results were "statistically accurate". For time-constrained healthcare contexts, all this suggests the need to identify other ways of establishing trust in the accuracy of AI models [68, 95]. Next, we synthesize and suggest strategies for establishing trust in AI applications for healthcare, and explain the tradeoffs we made in balancing the use of specific trust mechanisms with other requirements posed by the specific design context.

Balancing Sufficient Model Robustness with Clinical Utility: Amongst existing ideas and approaches to aid user trust in AI outputs are proposals to carefully consider when, and when *not*, to show predictions. In cases where prediction accuracy is lower and systems more likely to err, Jacobs et al. [52] suggest that predictions should perhaps not be shown altogether. Similar decisions were made by Beede et al. [9], who decided for their AI system that detects diabetic eye disease from retina images, to reject poorer quality images for analysis to reduce chances of incorrect assessments, even if the model could technically produce a strong prediction. Findings of a user study revealed how this created tensions among nurses, who reported frustration as they felt the images that they had taken as part of routine care, whilst human-readable, kept being rejected for AI analysis. Aside from considerations of technical robustness, Yang et al. [111] further proposed to only show AI prognosis in cases where there is "a meaningful disagreement" in clinician's assessment of the situation with the AI recommendation so as to minimize clinician burden; however identifying those instances of misalignment may prove challenging. In our work, we too deliberated choices about *limiting when predictions are shown in practice to try maximize the robustness, reliability and clinical usefulness of offered predictions*. This included: a prioritization of a very low false positive error rate (over false negatives); to only show predictions after three outcome measures, when they are more robust; and to only show predictions where clinically more relevant (i.e., by excluding predictions for patients with starting scores below RI thresholds, or below caseness). However, those restrictions also mean that predictions are not available earlier within treatment, where they could benefit especially those patients at risk of dropping out in the first 2–3 weeks of treatment; therefore presenting a tradeoff between maximizing for model robustness and clinical utility.

Engaging with Relevant Stakeholders and Demonstrating the Benefits of the AI: For establishing clinician trust in the accuracy of AI model outputs and thereby supporting the acceptability of innovative technology within healthcare [94], Sendak et al. [95] highlight the importance of engaging with target users in the design and development of the AI model and user interface. As part of those engagement the authors suggested to *demonstrate how the AI helps solve important problems for the specific users* (beyond technical innovation); and to *communicate the benefits of the AI application in ways that is directly relevant to those stakeholders*. It is indeed through our

engagement with iCBT supporters that we were able to develop a deeper understanding of their work practices and how AI could come to benefit them (Sections 3 and 5); and learn how to design and communicate AI outputs within the intervention (Section 5). We identified a number of additional insights:

1. Calibration and Fit with Existing Mental Models for Appropriately Interpreting Data Insights: In our work, we observed a certain *pragmatism* in how supporters evaluated issues of trust and the impact of prediction errors. Research by Cai et al. [16] too showed participants implicitly or explicitly describing how no AI tool (or person) is perfect. Similarly, when the supporters in our study reflected on the consequences of false predictions, they described the possibility of making mistakes in assessing a patients' situation not as something that is newly introduced by the AI, but as something that exists in their current work as well. Thus, supporters would instead *assess the benefits of having RI predictions available to them as a way to reduce uncertainty and errors in their own judgment.* Simultaneously, they were mindful that the AI insights would offer only one piece of information to their clinical assessment, and that this information comes with its own limitations—like any other data tools and measures. In other words, supporters arrived at this more adjusted, pragmatic understanding of the AI-derived data insight *through a comparison with other existing data practices and associated risk mitigation strategies.* For example, in interpretations of patient's clinical scores (PHQ-9 and GAD-7), supporters also described the need for caution in interpretations, as those scores may not exactly reflect what is going on for the person. While such clinical assessments can provide a numerical indicator and trend of that person's mental health progression that can be informative to clinical practice, they noted that these scores should not be given too much weight in isolation; and instead be assessed in the context of other information provided in patient messages and conversations. This feedback suggests their treating and evaluating of the AI output as a piece of information with similar limitations as these clinical scores. Nonetheless, what our research findings also highlighted is the importance to remind supporters of the need to balance assessments of the AI output in the context of other patient information to reduce data over-reliance. As this can be more complicated in time-constrained contexts, it also emphasizes the need for *careful staff training prior to any AI deployment to help ensure appropriate understanding and use* (see also [16]).

2. Balancing Human-AI Interactivity and Interrogations for Trust with Costs of Time and Interference: As mentioned above, we considered the AI predictions as a simple data insight that would primarily serve as a useful "signal" to aid prioritization of patient cases, but otherwise would not take away additional supporter time (i.e., to review explanations of the model output) to respect already tight review schedules. Other works on CDS tools, however, have indeed demonstrated how mechanisms such as: *visualizing the most important model parameters* [68]; and *offering refinement tools that allow clinicians to fine-tune or otherwise experiment with AI input parameters to alter algorithmic outputs* [15] can play a key role in supporting user understanding of the AI and its capabilities, and promote both AI transparency and trust. We too believe that showing, i.e., the most predictive feature(s) can potentially add useful insights above and beyond a binary prediction. We also saw in supporter's evaluation of the Population Comparison (Table 1, concept 4, and Appendix C) that additional explanations of the model can positively contribute to assessments of AI output credibility. Nonetheless, we were mindful that *the provision of additional information or other interactivity may also lead to supporters potentially reading too much into the AI result and consuming more of their review time*—especially in cases where additional data invites more ambiguity in the interpretation of the findings (cf. findings in Section 5.3.1). Worries about going too deep down into a thought process and tangential rabbit hole in reviewing and interacting with AI insights also surfaced in other related work [15]. Thus, aiming to take a first step towards introducing AI into this specific iCBT context, we would prioritize a simpler, easy-to-comprehend

AI insight that could "flag-up" if a patient was at risk of not fully benefiting from treatment, but otherwise would not take supporters too far away from their own thought processes and focus on the patient. With an increase in familiarity and understanding of AI use within clinical care, future work will likely expand on the scope and types of AI insights.

3. AI Credibility through Trusted Data Sources and Experiences of (Continued) Use: In the on-boarding of iCBT supporters to our design research, we found that explanations of our algorithms source of ground truth—the volume and type of data that the RI predictions are based on—contributed to their trust in the model outputs. As reported in our findings (Section 5.2.2), supporters would explicitly remark on the fact that the predictions were based on thousands of previous SilverCloud users. The large scale of the data (>46 K patient cases); its direct mapping to the specific application program; and its sourcing from the very company that the supporters work with and trust, impacted perceptions of the models' credibility and shaped supporters' assessment of the resulting AI output as potentially a "more reliable", "more objective, evidence-based" indicator than their own judgment. In keeping with other recent recommendation's for onboarding health professionals to AI system [16], this suggests the importance for key design decisions about data collection; source of ground truth; and model objectives to be made transparent to target users, both in prior-use training and interface design to aid transparency and the development of an appropriate mental model of the AI. Future work will also need to assess how trust in the predictions may develop and become calibrated through continuous use (cf. [16, 68]). In addition, for any real-world deployment, it is important to put measures into place to continuously oversee and closely monitor the on-going performance and reliability of the AI in-use [95] such that good performance and trust in the outputs are maintained over time.

4. Trust through Clinical Validation and External Approvals: Finally, there have been literature reports [16, 52, 68, 95, 111] that describe the need for *formal, internal and/ or external, rigorous clinical assessments of AI model validity or rather utility*, often in form of evidence-based methods (i.e., randomized controlled trials); research publications in prestigious journals; or FDA approval for health professionals' to be able to trust AI outputs. Especially for early-stage AI research and development, this highlights the importance to set and communicate appropriate expectations of what AI models can realistically achieve to-date and at various development stages to not prematurely diminish AI credibility [100]. Likely, gradually developing the AI, and building-up clinician's understanding of its workings and limitations—including demonstrations of technical validity and clinical utility—will also require the formation of longer-term healthcare partnerships.

6.3 Limitations and Future Work

Amongst the very many complex sociotechnical challenges involved in paving the way towards the successful development and adoption of AI interventions within real-world (mental) healthcare practices, our research focused on two specific aspects: the *identification* and *appropriate design* of a clinically useful AI application.

While our user research brought forward a wide range of possible and perhaps more impactful applications of AI in this context (see Appendix A and [99]), we chose to pursue RI predictions based on patient's frequent report of clinical scores. This was the outcome of a rather complex and lengthy design and development process that intersected multiple research fields to create an AI application that is clinically useful, practically feasible, and implementable within existing care. Above, we discussed the various tradeoffs we made to bring together: insights from user research; the clinical literature; and needs to achieve robust AI models from the data available with our goals to pursue a very human-centered and careful approach towards AI integration within an actual mental health context. We also acknowledge that our proposed user interface for RI prediction presents the result of multiple design tradeoffs and may not present an optimal data representation—especially with regards to concerns about over-trusting or contesting the AI

output, and for guiding supporters next steps towards specific actions. There further remains an open question about our focus on RI as the chosen outcome metric. It could be argued that a "RI" might present too high a hurdle for supporters to achieve, which can suggest improvement alone, without meeting the threshold of a significant change, could be portrayed as a negative.

Moreover, so far, our research only included the perspectives of a small number of iCBT supporters, who worked as PWPs at one specific NHS Trust in the UK. Not only do they present a rather homogenous group of low-intensity intervention specialists, their self-selection to engage in our research to investigate innovative AI uses may have also introduced a positivity bias towards the introduction of any such technology. In future work, we suggest a broader engagement with other stakeholders (i.e., different treatment localities) as well as patients.

As a next step towards addressing some of these issues, future research will need to: (i) delve deeper into clinician's experience with, and potential acceptance of, the developed RI prediction models (e.g., how access to the predictions may shape supporter practices, their sense of agency), (ii) investigate the effectiveness of their deployment for improving patient symptoms of depression and anxiety; and (iii) how these outcomes may differ for iCBT supporters with varying levels of expertise (novices vs. more experienced PWPs). To this end, and separate to the work presented here, researchers at SilverCloud Health have designed a large-scale randomized controlled trial (RCT)[7] to deepen understandings of the opportunities and unique challenges for how AI insights could come to support real-world healthcare practices, and benefit (mental) health patients.

7 CONCLUSION

Aiming to move beyond a focus on technical feasibility and advancing state-of-the-art algorithms that often regard the AI in isolation from their proposed use context, our research explored some of the unique challenges for designing and implementing an AI application for assisting the work practices of iCBT supporters. Specifically, we reported our iterative, human-centered approach to designing an AI application that predicts RI outcomes for patients receiving human-supported iCBT for depression and anxiety. Intersecting the fields of HCI, ML, and healthcare, we described how we engaged with iCBT supporters in interview and design research to (1) identify meaningful use scenarios and development targets for AI in this context; and (2) learn about the opportunities and challenges in designing and integrating specific AI insights within an established digital mental health service.

Our findings (i) provided key insights into the specific work practices and information needs of iCBT supporters, and, in response, (ii) outlined various opportunities for applications of AI in this context. Focusing on how AI could assist iCBT supporter's understanding of patient mental health progress to better identify possible barriers to improvement and engage in timely treatment adaptations, we (iii) explained our rationale for developing outcome prediction models. Specifically, we detailed our choices for the AI to contribute to FIT; predict RI; and ensure model robustness through adjustments for low false positive error rates, and predictions after at least three clinical measures. Further, we detailed design research that (iv) helped clarify use scenarios for the prediction models. Predominantly, iCBT supporters identified their potential use and utility as an "early indicator" for adapting patient treatment for improved outcomes, and in performance monitoring. This work also (v) brought forward multiple concerns about the use of outcome prediction in this specific mental health context, which pertained to supporter motivation to take action and performance pressures; potential uncritical treatment of, and over-reliance on, the prediction outcome that can cause reduced support and lesser engagement with, important individualized, patient-centric care; as well as reflections on the implications of "false" predictions for patients, supporters and care services. Our analysis (vi) concluded with

[7]For details see https://www.isrctn.com/ISRCTN18059067.

specific design implications for integrating AI models within a user interface, which marks an essential step forward towards real-world deployment.

In discussing what constitutes a human-centered approach to AI design in this healthcare context, we shared our perspective to focus on developing human-AI partnerships that seek to empower, not replace clinicians with AI by building on their professional expertise and protecting their sense of agency; and that give particular consideration to the therapeutic relationship between clinicians and patients, especially in digitally delivered interventions. For time-constrained healthcare contexts, we closed with reflections on how trust in AI applications for healthcare will need to become negotiated differently to commonly suggested (XAI) approaches of providing interpretable explanations of model outputs, and communications of statistical accuracy measures.

A APPENDIX

Additional Details on PWP Interview Study 1.

Study Purpose

The aim of this study was to: (i) better understand the specific information needs and challenges that iCBT supporters encounter for building up an understanding of their patients' progress and providing effective, personalized feedback; and, responding to these learnings, (ii) scope out potential opportunities for AI to help derive and support identified data review requirements.

Participants

We recruited 15 PWPs from the Berkshire Healthcare NHS Foundation Trust in the UK. All participants regularly acted as SilverCloud supporters as part of their role. Our interview sample predominantly self-reported as female (1 male) and generally included fully certified PWPs who were very experienced at using SilverCloud to support their patients (Table 3). Asked to rate their own level of experience as a SilverCloud supporter, the majority (n = 12) self-identified as "intermediate/ expert" or "expert". Participants also described having used the service a minimum of 1–2 years; some even reporting more than 5 years of use. Table 3 further shows the case load of each PWP at the time of interview, which most often included 15–30, or more SilverCloud patients. For this particular NHS service, these patient numbers typically represent ~50% of a supporter's overall case load.

Table 3. Participants Role Title and Level of Expertise in Supporting Patients via SilverCloud; the Amount of Years They Have Been Using the Service, and the Number of Patients Currently Assigned to Them.

Current Role	Supporter expertise	Years using SilverCloud	Patient load
Trainee PWP	Intermediate/Expert	1–2	4–5
Trainee PWP	Intermediate/Expert	1–2	30 or more
PWP	Intermediate	1–2	15–30
PWP	Intermediate	2–4	30 or more
PWP	Intermediate/Expert	2–4	15–30
PWP	Novice	1–2	15–30
PWP	Intermediate/Expert	1–2	30 or more
Senior PWP	Intermediate/Expert	4–5	5–15
Senior PWP	Intermediate/Expert	4–5	15–30
PWP Clinical Lead	Intermediate/Expert	2–4	15–30
PWP Clinical Lead	Intermediate/Expert	2–4	5–15
PWP Team Lead	Intermediate/Expert	5 or more	5–15
PWP Team Lead	Expert	4–5	15–30
PWP Team Lead	Intermediate/Expert	5 or more	4–5
Innovation & Service Improvement within Trust	Expert	5 or more	4–5

Procedure

All interviews were conducted remotely via video conferencing software (Microsoft Teams), and in one case via a phone call. Each interview lasted 1 hour and was audio recorded. The full interview guide can be accessed here [99]. All participants received a £30 gift voucher to a retail store of their choice to compensate them for their time spent in contributing to the research. The research study was carefully reviewed and monitored for compliance and privacy regulations; and approved by the NHS Health Research Authority (HRA, reference: 19/LO/1525). Each participant has been given a unique identification number to protect their anonymity.

Amongst others, we asked how these supporters currently identify whether their patients encounter any difficulties with the treatment program; which types of information they consider as most important in assessing their patients' situation; potential areas of uncertainty or knowledge gaps; and what challenges they might encounter when choosing what therapy contents or activities to recommend (see [99] for the interview guide).

Data Analysis

All audio recorded interviews were fully transcribed and subjected to Thematic Analysis [12]. This involved an intensive familiarization with the data, and the identification of, and system search for, reoccurring themes in the data that were developed in higher-level categories. Our analysis was guided by our main research question: What are the strategies and challenges of iCBT supporters for providing effective, personalized feedback to their mental health patients? Through this analysis, we identified six key themes that describe the importance, strategies, and challenges of iCBT supporters for providing effective, personalized feedback to patients. The themes are summarized below. A more detailed report is available here [99].

Main Findings

Summary of the six key themes that emerged from our thematic analysis about the tasks and challenges that PWPs encounter in providing effective, personalized iCBT feedback; as well as an outline of some of the questions and opportunities that AI research could explore in response.

Key Themes	AI Opportunities
(1) Understanding patients' mental health problems; risks; progress; and associated barriers for improvement that may need to be addressed	

A key focus in the review is for PWPs to gain a sufficient understanding of their patients' mental health. PWPs:
 	
(i) assess the patient's *main mental health problem* to be able to *identify the right treatment*, and *effectively focus* or *adapt the treatment* to the person's needs (i.e., through switching the patient to a different treatment program; extending the number or frequency of reviews; or step-up in care decisions);
 	
and (ii) regularly review the patients clinical scores and their changes over time to detect any *mental health risks*; potential *barriers to improvement* (i.e., persistent sleep difficulties) that should be addressed; or whether there are any *indicators of risk* that have to be responded to immediately.
	• How could AI support early or automatic detection of mental health risks?
	• How could AI assist PWPs' understanding of why a patient might not be improving in their mental health?
	  ○ How could AI support the identification of 'barriers to patient improvement' based on patients' clinical scores or specific markers in their language (i.e., in patient messages, text entries)?
	  ○ How could AI be leveraged to learn what mental health struggles (i.e., indicated via questionnaire items) may be particularly salient in predicting good or poor mental health outcomes?

Key Themes	AI Opportunities
Further, assessments of patient mental health progression and decisions to guide treatment are not solely done by considering a person's clinical scores, but (iii) *carefully evaluated in the context of other information* that the PWP might have about a patient and their life circumstances; and these are regularly discussed with other clinicians as part of CMS.	• How could AI help identify 'early' if a patient is likely to benefit from the chosen treatment, or not, to facilitate treatment adjustments and ensure patients get the 'right care'?
(2) Assessing and responding to patient engagement with iCBT program to provide targeted support	

PWPs identify specific *patterns in patient engagement* with the treatment; and review those behaviors in light of the patients' mental health progress.
 	
Responding to these patterns, PWPs tailor feedback to either (i) reinforce good engagement, or (ii) encourage more extensive or more frequent uses of the program, its tools, or the sharing of patient experiences and struggles to increase overall patient engagement, and to reduce risks of treatment drop-out.
 	
To better assess patient engagement, PWPs expressed (iii) the *desire to access duration information; and easier means to review patient engagement over time*. Temporal information and comparisons could help approximate how deeply a patient may review contents and support the noticing of changes or struggles in patient activity.
 	
Furthermore, PWPs explained the (iv) *limitations of providing effective feedback in cases where the person 'did not attend' (DNA) the iCBT program* at all during a review period.
	• How could AI support PWPs understanding of patient engagement?
	  ○ How could AI assist in the identification of 'engagement trends' or 'changes in behavioral patterns' that may be reflective of specific struggles?
	  ○ What are opportunities for employing AI to learn about characteristics of good/effective program uses vs. less beneficial uses?
	  ○ What (re-)engagement strategies might be most effective for different types of program engagers?
	• How could AI aid the earlier detection of a person's risk of 1st/2nd DNA (in the next review period) such that PWPs could reach out to patients earlier to support engagement and reduce risks of drop-out?

(3) Evaluating patient progress with the program to decide next treatment steps, and identifying program content struggles that require more explanation	
PWPs draw on available patient data to evaluate (i) *how well the person understands and utilizes the program contents and tools*. Most insight is derived from patient messages that explicitly communicate about learning progress or struggles; and patients' language use that can be indicative of their mental health state and behavior changes (i.e., indicators of positive change). Particular attention is given to the review of patients' use of program tools, which are crucial for helping them apply the content learnings to their situation and support the practice of core CBT skills. Building on their therapy expertise and familiarity with the SilverCloud tools, PWPs (ii) *look for any errors or sub-optimal use of the tools that need addressing*. To (iii) *help overcome identified misunderstandings and barriers to content or tool use*, PWPs employ a number of feedback strategies, such as: normalizing; the provision of alternative explanations; and additional learning materials. To effectively advance a patient through their treatment journey, PWPs (iv) *make recommendations for specific tools or contents to review* next; as informed by: the main treatment goal; patient pace and success in progressing through the program; and PWPs experience in what program components are particularly important for specific mental health conditions. In some instances, PWPs expressed the (v) *desire for more opportunities to personalize the content order and pace of treatment* to achieve a better tailoring to each person's needs.	• How could AI (i.e., NLP) assist in detecting whether patients are demonstrating learning experiences and improvements to their mental health (i.e., from positive changes in their language)? • What types of treatment interactions invite positive changes for different patient types? Could insights be leveraged to help ensure patients experience improvements (earlier)? • How could AI support the identification of patient struggles (i.e., false completions) of program contents? • How could AI support the effective personalization of program contents to each person's needs? ○ How could data insights be leveraged to give PWPs an indicator of how beneficial it may be to recommend/unlock a specific treatment module at a certain moment in time in the patients' treatment journey?

Key Themes	AI Opportunities
(4) Extracting and responding to 'relevant' patient information under time constraints to maximize the benefits of support	
PWPs described challenges to (i) *gather or extract treatment relevant patient information* to provide constructive, helpful feedback due to a reliance on online data; remote, asynchronous communication; and working under time constraints. In cases where they have too little information about a patient (i.e., due to low engagement), this carries risks that PWPs may form potentially false assumptions about the patient. Further, therapeutic rapport can be harder to establish, which can reduce opportunities for patients to open-up. On the contrary, where too much data is available to review, it can take PWPs too long to filter through patient content. To address this, PWPs described strategies for: (ii) *inviting more patient input* by deliberately directing specific questions at them and actively inviting online messaging; and (iii) *soliciting key information* from patient contents, for example, by filtering the text for specific treatment relevant key words (i.e., medication; mood expressions, life circumstances). Further, to be able to (iv) *respond in a time-efficient manner,* PWPs make use of pre-written text paragraphs and message templates. Mostly, these are used for very standardized responses (i.e., DNA events), or otherwise they are carefully personalized to not be perceived as ridged or repetitive; and instead, to achieve the all-important person-centered feel.	• How could AI help extract/foreground relevant patient information (i.e., language indicators related to medication, treatment episodes)? • How could AI provide additional feedback to PWPs to aid their understanding of the patients main problem and what about the treatment is or isn't (potentially) working for them? • How could AI invite more opportunities for patient input/feedback and dialogue? For example, for a particular patient type, what communication strategies would make it more likely for them to respond to their supporter? • How could AI reduce time-consuming labor (e.g., auto-complete generic summaries) for example as part of template uses? How to translate individual templates as a machine-readable resource?
***(5) Gaining a 'sense of the person' and forming a therapeutic alliance* to achieve a more personal connection, for improved patient outcomes**	
PWP described the role and (i) *importance of 'personalized support' and establishing a therapeutic relationship with their patients* for improving patient engagement and self-disclosure. Reported key benefits include: reduced perceptions of support being delivered through a machine rather than a human being who listens, understands, and cares for the person; improved patient trust and hope in the effectiveness of an 'online' approach; assurance that the same standard and quality of care is delivered via online treatment as is in other therapy formats; as well as better patient engagement with the program and responsiveness to their PWP. All these factors promote overall therapy outcomes. Outlining (ii) *strategies for providing personalized support,* PWPs described explicit efforts to convey in their communication a sense of them being a real person and ensure that the patient feels heard, understood, and cared for. Amongst others they reported: the use of person-identifiers such as their own and other peoples' names; choices in the timing and writing style of their messages; asking the patient personally relevant questions; reflecting and referring back to them what they had said or done previously; recognizing and normalizing patient struggles; and to respond with empathy and compassion.	• How could AI enable PWPs to get a better 'sense of the person' (to care & connect)? • How could AI advance PWPs' understanding of what types of their communication strategies (i.e., encouragements, normalization, etc.) are most effective for a particular patient(type)? • How personal/personable (or frequent) should PWP feedback be to convey a sense of 'human care'?

Key Themes	AI Opportunities
(6) Developing PWP skills and confidence in effectively communicating with patients, especially where supporters are iCBT novices	
Describing key competencies and configurations of supporter feedback, PWPs highlighted how especially (i) *novice supporters encounter numerous difficulties in getting started with* digital therapy offerings like iCBT. It takes time for new PWPs to build-up the skills and professional confidence to be effective in this role that requires: familiarization with extensive and changing treatment programs; and the translation of 'common factor skills' into written communications as well as knowledge and experience of how to effectively personalize feedback messages. This raises questions of how future learning and training can best be supported. Lastly, supporters described the benefits and a desire for (ii) *receiving more feedback about how their actions impact patients*, which can help clarify reasons for patient drop-out, or improve their confidence in decisions around patient care. This can enhance treatment effectiveness and add to PWP confidence, and motivation and enjoyment of their work.	• How could AI assist especially novice supporters to communicate most effectively and build-up their confidence in feedback message writing (e.g., by providing smart guidance on "what to say" when/tailoring communication style)? • How could AI help supporters better understand how their actions matter: what types of actions (and their timing within the care pathway) achieve positive change/have the most impact?

B APPENDIX

RNN model architecture.

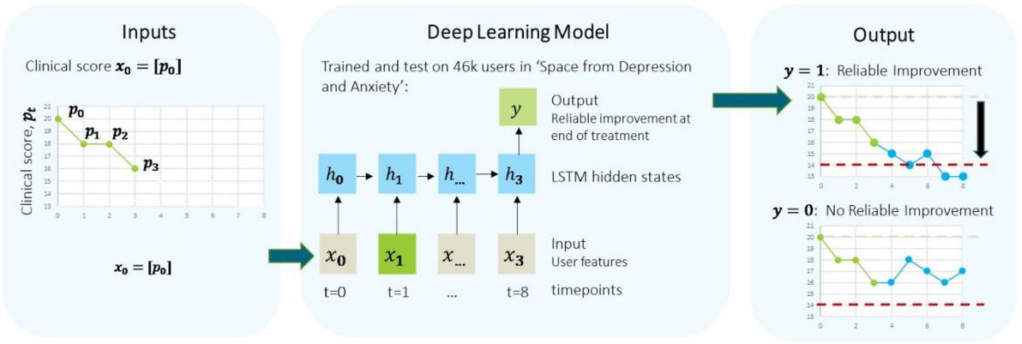


Fig. 5. The RNN architecture used to predict RI in depression or anxiety by the end of the treatment (typically an 8 week/review period). We used early clinical scores in PHQ-9 and GAD-7 for depression and anxiety respectively as the input features. The data is split into 70:20:10 for training, validation, and testing. Both trained models are 3-layer RNNs with a 50-dimensional hidden layer with long-short term memory units (LSTM) to encode patient features at different time points throughout treatment. Model training and testing was done on the AzureML infrastructure, in Python 3.7; and all RNNs were implemented using PyTorch. For more detail on the development and performance of these models see Prasad et al. [82].

C APPENDIX

Design feedback on the six prediction outcome concepts.

Wheel Indicator	The Wheel Indicator was assessed as a 'very visual' concept that enables easy, near instant understanding due to the graded chart with its clear green and red areas; providing a lot of information at a glance. While one supporter suggested its size was too big, others liked its proportions and combination with large numbers as it meant the visual cannot be overlooked. Many supporters explicitly expressed their liking of the confidence percentages, suggesting that it provides more detail than a simpler yes-no indication, whilst others raised concerns about how these numbers are to be interpreted. Some supporters described to potentially get hung up on the detail of the numbers, and to not have time to reflect about them under time pressure. Further criticized was potential confusion that the Wheel indicator could be read as the spectrum from deterioration (far left) to improvement (far right) rather than as a binary indicator for RI (no vs. yes). One of the supporters also had a very strong reaction to the symbolism of the wheel and 'red NO' label that to them felt like a very 'stark', 'attacking' indicator alike a warning signal in cars that 'screams failure' or 'danger', which raised concerns about the visual potentially feeling 'really demoralizing' to, and evoking stress in, PWPs.
Multiple Outcomes Table	The Multiple Outcomes table concept received much criticism. Compared to other designs, it was assessed as least visually pleasing, and felt 'too wordy' and 'too complex' to understand at a glance. It takes time to fully understand, not confuse, and get used to the spectrum of definitions and outputs shown in the table; suggesting a more slimline approach (i.e., only show most important definition and others on demand; show results for PHQ and GAD in separate tables). Yet, once the initial learning hurdle is taken, supporters described that having access to multiple definitions can add value as these all present all the types of information reviewed during line management and when patients are discharged. While few PWPs valued illustrations of the prediction confidence through the color gradient over text, others thought the color coding could be clearer or easier to learn. Similarly contested was the clarity of text labels used, which the majority of supporters found 'too vague'. They struggled to understand the nuances or categories that distinguish 'very likely' from 'likely' labels', which often led them to express a clear preference for, and proposals to add, numerical values.
Visual Cue + Text	The Visual Cue + Text concept was received best. It was assessed as a fairly 'neutral' indicator with a clear, succinct and easy-to-access presentation that would convey information at a quick glance. The 'little grading scale with the little arrow on' was found to aid interpretations of the prediction results alongside the text labels. While for some of the supporters the visual cue and text representation by themselves were sufficient, others thought the text labels were too ambiguous, and assessed the visual cue as 'too small'. To improve the graphic, supporters described that a larger graphic could facilitate pinpointing the 'whereabouts' of the arrows for PHQ vs. GAD; and proposed to add percentage numbers to cater for the variance in information representation preferences amongst the supporters.

Population Comparison	Although the Population Comparison combines many of the other design components including bar charts and percentage numbers, all supporters unanimously describe the concept as 'too detailed and complicated'. Their busy schedules mean they do not have the time needed to process and comprehend all the information offered. For some, having the RI prediction shown on a 0-100 continuum (rather than a yes/no dial) also added confusion about how the numbers are to be interpreted. Despite these concerns, they thought the information offered through the option to compare the patient with 'the population, or other people who had similar presentations' to: be generally really helpful; offer a useful context for interpretations of their treatment choices (i.e., a less than average chance of RI will influence how closely to monitor/support patients); and usefully ground the prediction within real-world data by providing an explanation that can help in understanding the AI prediction, and user trust. While some supporters also saw the benefits of being able to give oneself a 'pat on the back' if the visual showed for their own patient a higher chance of RI than for the general population; this was a concern for others, who found comparisons could add to supporter stress and competition. With its proposed focus on comparing the individual to a group of similar patients, the relevance of this design for facilitating individualized, person-centered care or CMS meetings remained unclear. Many supporters therefore considered this design concept more as an add-on, or optional information resource.
Client Dashboard	For the Client Dashboard, the majority of supporters thought the prediction data was 'nicely laid out' within the table and easy-to-understand due to the use of clear percentages. Few thought the many numbers were overwhelming and could get accidentally confused. As the natural entry point to their patient review practices, the predictions were easy-to-locate and 'straight away accessible', which also meant, however, that they could be 'easily skipped', especially since the dashboard is mostly used to navigate to a patient's profile, and not to access or compare patient scores. Nonetheless, the dashboard was assessed to provide a useful overview across patients and their RI prognosis to prime PWPs prior to a review about what to expect from the session and guide resource allocation, time planning, or discharge decisions. The dashboard also enables the spotting of trends in supporter performance across patients, which some supervisors regarded as useful insight, while other PWPs raised concerns about it potentially reflecting badly on their work practices, especially if many patients are predicted as unlikely to improve. What might complicate these reviews is that the predictions shown reflect patient outcomes at different stages in their treatment, which may be less suitable or appropriate to compare across. Thus, overall, this concept was regarded as an 'optional' add-on that could provide additional context to CMS reviews, but may be less relevant in supporting important individualized care.

ACKNOWLEDGEMENTS

Special thanks go to all our PWP participants for their help and input to the research; as well as the reviewers of this article for their valuable suggestions and thoughtful feedback on this work. We further thank our colleagues and collaborators for all their contributions: Junaid Bajwa, Danielle Belgrave, James Bligh, Isabel Chien, Catalina Cumpanasoiu, Caroline Earley, Dessie Keegan, Usman Munir, Hannah Murfet, Aditya Nori, David O'Callaghan, Niranjani Prasad, Tim Regan, Derek Richards, Angel E. Roig and Ryutaro Tanno.

REFERENCES

[1] Marios Adamou, Grigoris Antoniou, Elissavet Greasidou, Vincenzo Lagani, Paulos Charonyktakis, and Ioannis Tsamardinos. 2018. Mining free-text medical notes for suicide risk assessment. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM, 8 pages. DOI: https://doi.org/10.1145/3200947.3201020

[2] Jesus S. Aguilar-Ruiz, Raquel Costa, and Federico Divina. 2004. Knowledge discovery from doctor–patient relationship. In *Proceedings of the 2004 ACM Symposium on Applied Computing*. ACM, 280–284. DOI: https://doi.org/10.1145/967900.967960

[3] Md. Golam Rabiul Alam, Eung Jun Cho, Eui-Nam Huh, and Choong Seon Hong. 2014. Cloud based mental state monitoring system for suicide risk reconnaissance using wearable bio-sensors. In *Proceedings of the 8th International*

Conference on Ubiquitous Information Management and Communication. ACM, 6 pages. DOI: https://doi.org/10.1145/2557977.2558020

[4] Shehzad Ali, Laura Rhodes, Omar Moreea, Dean McMillan, Simon Gilbody, Chris Leach, Mike Lucock, Wolfgang Lutz, and Jaime Delgadillo. 2017. How durable is the effect of low intensity CBT for depression and anxiety? Remission and relapse in a longitudinal cohort study. *Behaviour Research and Therapy* 94 (2017), 1–8. DOI: https://doi.org/10.1016/j.brat.2017.04.006

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. DOI: https://doi.org/10.1145/3290605.3300233

[6] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–5. DOI: https://doi.org/10.1145/3411763.3441347

[7] Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* 145, PT B (2017), 137–165. DOI: https://doi.org/10.1016/j.neuroimage.2016.02.079

[8] A. B. M. Asadullah, Alim Al Ayub Ahmed, and Praveen Kumar Donepudi. 2020. Artificial intelligence in clinical genomics and healthcare. *European Journal of Molecular and Clinical Medicine* 7, 11 (2020), 1194–1202.

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–12. DOI: https://doi.org/10.1145/3313831.3376718

[10] Shadi Beshai, Keith S. Dobson, Claudi LH Bockting, and Leanne Quigley. 2011. Relapse and recurrence prevention in depression: Current research and future prospects. *Clinical Psychology Review* 31, 8 (2011), 1349–1360. DOI: https://doi.org/10.1016/j.cpr.2011.09.003

[11] Claire Bone, Melanie Simmonds-Buckley, Richard Thwaites, David Sandford, Mariia Merzhvynska, Julian Rubel, Anne-Katharina Deisenhofer, Wolfgang Lutz, and Jaime Delgadillo. 2021. Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health* 3, 4 (2021), e231–e240. DOI: https://doi.org/10.1016/S2589-7500(21)00018-2

[12] Virginia Braun, and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[13] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 454–464. DOI: https://doi.org/10.1145/3377325.3377498

[14] Danilo Bzdok, and Andreas Meyer-Lindenberg. 2018. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, 3 (2018), 223–230. DOI: https://doi.org/10.1016/j.bpsc.2017.11.007

[15] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–14. DOI: https://doi.org/10.1145/3290605.3300234

[16] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI Collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 24 pages. DOI: https://doi.org/10.1145/3359206

[17] Xiongcai Cai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. 2016. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association* 23, 3 (2016), 553–561. DOI: https://doi.org/10.1093/jamia/ocv110

[18] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S. Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D. Leow. 2017. DeepMood: Modeling mobile phone typing dynamics for mood detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 747–755. DOI: https://doi.org/10.1145/3097983.3098086

[19] Stevie Chancellor. 2018. Computational methods to understand deviant mental wellness communities. In *Proceedings of the Extended Abstracts CHI 2018. ACM, Paper DC05*. DOI: https://doi.org/10.1145/3170427.3173021

[20] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered machine learning: the case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 32 pages. DOI: https://doi.org/10.1145/3359249

[21] Adam M. Chekroud, Julia Bondar, Jaime Delgadillo, Gavin Doherty, Akash Wasil, Marjolein Fokkema, Zachary Cohen, Danielle Belgrave, Robert DeRubeis, Raquel Iniesta, Dominic Dwyer, and Karmel Choi. 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 20, 2 (2021), 154-170. DOI: https://doi.org/10.1002/wps.20882

[22] Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Proceedings of the Companion the Web Conference 2018*, 1653–1660. DOI: https://doi.org/10.1145/3184558.3191624

[23] Isabel Chien, Angel Enrique, Jorge Palacios, Tim Regan, Dessie Keegan, David Carter, Sebastian Tschiatschek, Aditya Nori, Anja Thieme, Derek Richards, Gavin Doherty, and Danielle Belgrave. 2020. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. *JAMA Network Open* 3, 7 (2020), e2010791–e2010791. DOI: https://doi.org/10.1001/jamanetworkopen.2020.10791

[24] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–16. DOI: https://doi.org/10.1145/3313831.3376341

[25] David M. Clark. 2021. *Improving Access to Psychological Therapies Manual*. London: NHS England. 2021. Retrieved from https://www.england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual-v5.pdf.

[26] Enrico Coiera. 2019. The last mile: where artificial intelligence meets reality. *Journal of Medical Internet Research* 21, 11 (2019), e16323. DOI: https://doi.org/10.2196/16323

[27] Jaime Delgadillo, Karen Overend, Mike Lucock, Martin Groom, Naomi Kirby, Dean McMillan, Simon Gilbody, Wolfgang Lutz, Julian A. Rubel, and Kim de Jong. 2017. Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy* 99 (2017), 89–97. DOI: https://doi.org/10.1016/j.brat.2017.09.011

[28] Jaime Delgadillo, Kim de Jong, Mike Lucock, Wolfgang Lutz, Julian Rubel, Simon Gilbody, Shehzad Ali, Elisa Aguirre, Mark Appleton, Jacqueline Nevin, Harry O'Hayon, Ushma Patel, Andrew Sainty, Peter Spencer, and Dean McMillan. 2018. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry* 5, 7 (2018), 564–572. DOI: https://doi.org/10.1016/S2215-0366(18)30162-7

[29] Orianna DeMasi and Benjamin Recht. 2017. A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 763–771. DOI: https://doi.org/10.1145/3123024.3125609

[30] Srikant Devaraj, Sushil K. Sharma, Dyan J. Fausto, Sara Viernes, and Hadi Kharrazi. 2014. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research* 3, 2 (2014), 36–53. DOI: http://dx.doi.org/10.5430/jbar.v3n2p36

[31] Gavin Doherty, David Coyle, and John Sharry. 2012. Engagement with online mental health interventions: An exploratory clinical study of a treatment for depression. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1421–1430. DOI: https://doi.org/10.1145/2207676.2208602

[32] Afsaneh Doryab, Mads Frost, Maria Faurholt-Jepsen, Lars V. Kessing, and Jakob E. Bardram. 2015. Impact factor analysis: Combining prediction with parameter ranking to reveal the impact of behavior on health outcome. *Personal Ubiquitous Computing* 19, 2 (2015), 355–365. DOI: http://dx.doi.org/10.1007/s00779-014-0826-8

[33] Karin Drivenes, Vegard Øksendal Haaland, Terje Mesel, and Lars Tanum. 2019. Practitioners' positive attitudes promote shared decision-making in mental health care. *Journal of Evaluation in Clinical Practice* 25, 6 (2019), 1041–1049. DOI: https://doi.org/10.1111/jep.13275

[34] Daniel Duffy, Angel Enrique, Sarah Connell, Conor Connolly, and Derek Richards. 2020. Internet-delivered cognitive behavior therapy as a prequel to face-to-face therapy for depression and anxiety: a naturalistic observation. *Frontiers in Psychiatry* 10 (2020), 902. DOI: https://doi.org/10.3389/fpsyt.2019.00902

[35] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. ACM, 211–223. DOI: https://doi.org/10.1145/3172944.3172961

[36] Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 16 pages. DOI: https://doi.org/10.1145/3290605.3300364

[37] Chaonan Feng, Huimin Gao, Xuefeng B. Ling, Jun Ji, and Yantao Ma. 2018. Shorten bipolarity checklist for the differentiation of subtypes of bipolar disorder using machine learning. In *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*. ACM, 162–166. DOI: https://doi.org/10.1145/3194480.3194508

[38] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health* 4, 2 (2017), e7785. DOI: https://doi.org/10.2196/mental.7785

[39] Erik Forsell, Erik, Nils Isacsson, Kerstin Blom, Susanna Jernelöv, Fehmi Ben Abdesslem, Nils Lindefors, Magnus Boman, and Viktor Kaldo. 2019. Predicting treatment failure in regular care Internet-Delivered Cognitive Behavior Therapy for depression and anxiety using only weekly symptom measures. *Journal of Consulting and Clinical Psychology* 88, 4 (2019), 311–321. DOI: https://doi.org/10.1037/ccp0000462

[40] Erik Forsell, Susanna Jernelöv, Kerstin Blom, Martin Kraepelien, Cecilia Svanborg, Gerhard Andersson, Nils Lindefors, and Viktor Kaldo. 2019. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: A single-blind randomized clinical trial with insomnia patients. *American Journal of Psychiatry* 176, 4 (2019), 315–323. DOI: https://doi.org/10.1176/appi.ajp.2018.18060699

[41] Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen, and Michael Alexander Riegler. 2019. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In *Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care*. ACM, 9–15. DOI: https://doi.org/10.1145/3347444.3356238

[42] Dimitrios Galiatsatos, Georgia Konstantopoulou, George Anastassopoulos, Marina Nerantzaki, Konstantinos Assimakopoulos, and Dimitrios Lymberopoulos. 2015. Classification of the most significant psychological symptoms in mental patients with depression using bayesian network. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks*. Lazaros Iliadis and Chrisina Jane (Eds.), ACM, 8 pages. DOI: https://doi.org/10.1145/2797143.2797159

[43] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "Let me tell you about your mental health!": Contextualized classification of reddit posts to DSM-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 753–762. DOI: https://doi.org/10.1145/3269206.3271732

[44] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Joint Summits on Translational Science Proceedings*, AMIA Joint Summits on Translational Science, 191–200. DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/

[45] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous stress detection using a wrist device: In laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: AdjunctAdjunct*, 1185–1193. DOI: https://doi.org/10.1145/2968219.2968306

[46] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian G. K. Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L. Lewis, Richard M. Wexler, and Dominick L. Frosch. 2013. "Many miles to go…": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Medical Informatics and Decision Making* 13, 2 (2013), 1–10. DOI: https://doi.org/10.1186/1472-6947-13-S2-S14

[47] Kevin Hilbert, Stefanie L. Kunas, Ulrike Lueken, Norbert Kathmann, Thomas Fydrich, and Lydia Fehm. 2020. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. *Behaviour Research and Therapy* 124 (2020), 103530. DOI: http://dx.doi.org/10.1016/j.brat.2019.103530

[48] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 95–99. DOI: https://doi.org/10.1145/3064663.3064703

[49] Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D. Caperton, David C. Atkins, and Zac E. Imel. 2018. "It's hard to argue with a computer": Investigating Psychotherapists' Attitudes towards Automated Evaluation. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 559–571. DOI: https://doi.org/10.1145/3196709.3196776

[50] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. DOI: https://doi.org/10.1145/3290605.3300809

[51] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106. DOI: https://doi.org/10.2196/12106

[52] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–14. DOI: https://doi.org/10.1145/3411764.3445385

[53] Neil S. Jacobson, and Paula Truax. 1992. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. In *Proceedings of the Methodological Issues & Strategies in Clinical Research*. American Psychological Association, 631–648. DOI: https://doi.org/10.1037/10109-042

[54] Robert Johansson, Gerhard Andersson, Ebmeier, Smit, Kessler, Cuijpers, Cuijpers, Andersson, Andersson, Andersson, and others. 2012. Internet-based psychological treatments for depression. *Expert Review of Neurotherapeutics* 12, 7 (2012), 861–870. DOI: https://doi.org/10.1586/ern.12.63

[55] Kevin B. Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E. Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L. Snowdon. 2021. Precision medicine, AI, and the future of personalized health care. *Clinical and Translational Science* 14, 1 (2021), 86–93. DOI: https://doi.org/10.1111/cts.12884

[56] Michael I. Jordan and Tom M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. DOI: https://doi.org/10.1126/science.aaa8415

[57] Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. 2015. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2201–2210. DOI: https://doi.org/10.1145/2702123.2702479

[58] Ramakanth Kavuluru, María Ramos-Morales, Tara Holaday, Amanda G. Williams, Laura Haye, and Julie Cerel. 2016. Classification of helpful comments on online suicide watch forums. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 32–40. DOI: https://doi.org/10.1145/2975167.297517

[59] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. 2018. Can a machine tend to teenagers' Emotional Needs? A study with conversational agents. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–6. DOI: https://doi.org/10.1145/3170427.3188548

[60] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 9 (2001), 606–613. DOI: https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[61] Emily Lattie, Katherine A. Cohen, Nathan Winquist, and David C. Mohr. 2020. Examining an app-based mental health self-care program, intellicare for college students: single-arm pilot study. *JMIR Mental Health* 7, 10 (2020), e21075. DOI: https://doi.org/10.2196/21075

[62] Yena Lee, Renee-Marie Ragguett, Rodrigo B. Mansur, Justin J. Boutilier, Joshua D. Rosenblat, Alisson Trevizol, Elisa Brietzke, Kangguang Lin, Zihang Pan, Mehala Subramaniapillai, Timothy C. Y. Chan, Dominika Fus, Caroline Park, Natalie Musial, Hannah Zuckerman, Vincent Chin-Hung Chen, Roger Ho, Carola Rong, and Roger S. McIntyre. 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders* 241 (2018), 519–532. DOI: https://doi.org/10.1016/j.jad.2018.08.073

[63] Isaac M. Lipkus, and Justin G. Hollands. 1999. The visual communication of risk. *JNCI Monographs* 1999, 25 (1999), 149–163. DOI: https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191

[64] Bernd Löwe, Oliver Decker, Stefanie Müller, Elmar Brähler, Dieter Schellberg, Wolfgang Herzog, and Philipp Yorck Herzberg. 2008. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care* 46, 3 (2008), 266–274. https://www.jstor.org/stable/40221654.

[65] Lorenzo Lorenzo-Luaces, Robert J. DeRubeis, Annemieke van Straten, and Bea Tiemens. 2017. A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders* 213 (2017), 78–85. DOI: https://doi.org/10.1016/j.jad.2017.02.010

[66] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37, (2014), 94–100. DOI: https://doi.org/10.1016/j.chb.2014.04.043

[67] Adria Mallol-Ragolta, Svati Dhamija, and Terrance E. Boult. 2018. A multimodal approach for predicting changes in PTSD symptom severity. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, 324–333. DOI: https://doi.org/10.1145/3242969.3242981

[68] Matthiesen Stina, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T. Philbert, Jesper Hastrup Svendsen, and Tariq Osman Andersen. 2021. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: Near-Live feasibility and qualitative study. *JMIR Human Factors* 8, 4 (2021), e26964. DOI: https://doi.org/10.2196/26964

[69] Andrew M. McIntosh, Robert Stewart, Ann John, Daniel J. Smith, Katrina Davis, Cathie Sudlow, Aiden Corvin, Kristin K. Nicodemus, David Kingdon, Lamiece Hassan, Matthew Hotopf, Stephen M. Lawrie, Tom C. Russ, John R. Geddes, Miranda Wolpert, Eva Wölbert, and David J. Porteous. 2016. Data science for mental health: A UK perspective on a global challenge. *The Lancet Psychiatry* 3, 10 (2016), 993–998. DOI: https://doi.org/10.1016/S2215-0366(16)30089-X

[70] Adam S. Miner, Nigam Shah, Kim D. Bullock, Bruce A. Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key Considerations for Incorporating Conversational AI in Psychotherapy. *Front Psychiatry* 10, (2019). DOI: https://doi.org/10.3389/fpsyt.2019.00746

[71] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology* 13, 23–47. https://doi.org/10.1146%2Fannurev-clinpsy-032816-044949

[72] Thin Nguyen, Bridianne O'Dea, Mark Larsen, Dinh Phung, Svetha Venkatesh, and Helen Christensen. 2017. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications* 76, 8 (2017), 10653–10676. DOI: https://doi.org/10.1007/s11042-015-3128-x

[73] Alicia L. Nobles, Jeffrey J. Glenn, Kamran Kowsari, Bethany A. Teachman, and Laura E. Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 11 pages. DOI: https://doi.org/10.1145/3173574.3173987

[74] Ehimwenma Nosakhare and Rosalind Picard. 2019. Probabilistic latent variable modeling for assessing behavioral influences on well-being. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2718–2726. DOI: https://doi.org/10.1145/3292500.3330738

[75] Blessing Ojeme, and Audrey Mbogho. 2016. Selecting learning algorithms for simultaneous identification of depression and comorbid disorders. *Procedia Computer Science* 96, 1294–1303. DOI: https://doi.org/10.1016/j.procs.2016.08.174

[76] Ozan Oktay, Jay Nanavati, Anton Schwaighofer, David Carter, Melissa Bristow, Ryutaro Tanno, Rajesh Jena, Gill Barnett, David Noble, Yvonne Rimmer, Ben Glocker, Kenton O'Hara, Christopher Bishop, Javier Alvarez-Valle, and Aditya Nori. 2020. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. *JAMA Network Open* 3, 11 (2020), e2027426–e2027426. DOI: https://doi.org/10.1001/jamanetworkopen.2020.27426

[77] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with stress through pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. ICST (Institute for Computer Sciences, Social Informatics and Telecommunications Engineering)*, 109–117. DOI: http://dx.doi.org/10.4108/icst.pervasivehealth.2014.255070

[78] Albert Park, Mike Conway, and Annie T. Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: A text mining and visualization approach. *Computers in Human Behavior* 78 (2018), 98–112. DOI: https://doi.org/10.1016/j.chb.2017.09.001

[79] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Proceedings of the Designing Interactive Systems Conference 2021*. Association for Computing Machinery, 926–941. DOI: https://doi.org/10.1145/3461778.3462143

[80] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S. Lasecki. 2019. Identifying challenges and opportunities in Human-AI collaboration in healthcare. In *Proceedings of the Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 506–510. DOI: https://doi.org/10.1145/3311957.3359433

[81] Pawel Matykiewicz Pestian, and Jacqueline Grupp-Phelan. 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. ACM, 96–97.

[82] Niranjani Prasad, Isabel Chien, Tim Regan, Angel Enrique, Jorge Palacios, Dessie Keegan, Usman Munir, Ryutaro Tanno, Hannah Murfet, Aditya Nori, Derek Richards, Gavin Doherty, Danielle Belgrave, Anja Thieme. Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. *PLOS One* (submitted, under review), 1–31.

[83] Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of change: Analyzing Peer-Based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. DOI: https://doi.org/10.1145/3290605.3300294

[84] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. ACM, 385–394. DOI: https://doi.org/10.1145/2030112.2030164

[85] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 1-10. DOI: https://doi.org/10.1038/s41746-018-0029-1

[86] Derek Richards, Ladislav Timulak, Emma O'Brien, Claire Hayes, Noemi Vigano, John Sharry, and G. Doherty. 2015. A randomized controlled trial of an internet-delivered treatment: its potential as a low-intensity community intervention for adults with symptoms of depression. *Behaviour Research and Therapy* 75, 20–31. DOI: https://doi.org/10.1016/j.brat.2015.10.005

[87] Derek Richards, Angel Enrique, Nora Eilert, Matthew Franklin, Jorge Palacios, Daniel Duffy, Caroline Earley, Judith Chapman, Grace Jell, Sarah Sollesse, and Ladislav Timulak. 2020. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *NPJ Digital Medicine* 3, 85 (2020), 1–10. DOI: https://doi.org/10.1038/s41746-020-0293-8

[88] Darius A. Rohani, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2018. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptomsin patients with affective disorders: Systematic review. *JMIR mHealth and uHealth* 6, 8 (2018), e165. DOI: https://doi.org/10.2196/mhealth.9691

[89] Asif Salekin, Jeremy W. Eberle, Jeffrey J. Glenn, Bethany A. Teachman, and John A. Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on Interactive Mobile Wearable Ubiquitous Technology* 2, 2, (2018), 26 pages. DOI: https://doi.org/10.1145/3214284

[90] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. 2019. Communicating Uncertainty in Fertility Prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–11. DOI: https://doi.org/10.1145/3290605.3300391

[91] Stephen M. Schueller, Kathryn Noth Tomasino, and David C. Mohr. 2017. Integrating human support into behavioral intervention technologies: the efficiency model of support. *Clinical Psychology: Science and Practice* 24, 1 (2017), 27–45. DOI: https://psycnet.apa.org/doi/10.1037/h0101740

[92] Jennie Sharf, Louis H. Primavera, and Marc J. Diener. 2010. Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training* 47, 4 (2010), 637–645. DOI: https://psycnet.apa.org/doi/10.1037/a0021175

[93] Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine* 49, 9 (2019), 1426–1448. DOI: https://doi.org/10.1017/S0033291719000151

[94] Mandeep Sekhon, Martin Cartwright, and Jill J. Francis. 2017. Acceptability of healthcare interventions: An overview of reviews and development of a theoretical framework. *BMC Health Services Research* 17, 1 (2017), 1–13. DOI: https://doi.org/10.1186/s12913-017-2031-8

[95] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 99–109. DOI: https://doi.org/10.1145/3351095.3372827

[96] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Passive mobile sensing and psychological traits for large scale mood prediction. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 272–281. DOI: https://doi.org/10.1145/3329189.3329213

[97] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing uncertainty about the future. *Science* 333, 6048 (2011), 1393–1400. DOI: https://doi.org/10.1126/science.1191181

[98] M. Srividya, S. Mohanavalli, and N. Bhalaji. 2018. Behavioral modeling for mental health using machine learning algorithms. *Journal of Medical Systems* 42, 5 (2018), 88. DOI: https://doi.org/10.1007/s10916-018-0934-5

[99] Anja Thieme. 2021. Understanding the Information Needs and Practices of Human Supporters of an Online Mental Health Intervention to Inform Machine Learning Applications. arXiv:4006845. Retrieved from https://arxiv.org/abs/4006845.

[100] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction* 27, 5 (2020), 53 pages. DOI: https://doi.org/10.1145/3398069

[101] Anja Thieme, Ed Cutrell, Cecily Morrison, Alex Taylor, and Abigail Sellen. 2020. Interpretability as a dynamic of human-AI interaction. *Interactions* 27, 5 (2020), 40–45. DOI: https://doi.org/10.1145/3411286

[102] Truyen Tran, Dinh Phung, Wei Luo, Richard Harvey, Michael Berk, and Svetha Venkatesh. 2013. An integrated framework forsuicide risk prediction. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, and Jingrui He (Eds.), ACM, 1410–1418. DOI: https://doi.org/10.1145/2487575.2488196

[103] Ingrid S. van Maurik, Leonie N. C. Visser, Ruth E. Pel-Littel, Marieke M. van Buchem, Marissa D. Zwan, Marleen Kunneman, Wiesje Pelkmans, Femke H. Bouwman, Mirella Minkman, Niki Schoonenboom, Philip Scheltens, Ellen M. A. Smets, and Wiesje M. van der Flier. 2019. Development and usability of ADappt: web-based tool to support

clinicians, patients, and caregivers in the diagnosis of mild cognitive impairment and Alzheimer disease. *JMIR Formative Research* 3, 3 (2019), e13417. DOI: https://doi.org/10.2196/13417

[104] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in rural Clinics: Challenges in AI-Powered clinical decision support system deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–18. DOI: https://doi.org/10.1145/3411764.3445432

[105] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–15. DOI: https://doi.org/10.1145/3290605.3300831

[106] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14. DOI: https://doi.org/10.1145/2632048.2632054

[107] Harvey A. Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J. Baxter, Alize J. Ferrari, Holly E. Erskine, Fiona J. Charlson, Rosana E. Norman, Abraham D. Flaxman, Nicole Johns, Roy Burstein, Christopher J. L. Murray, and Theo Vos. 2013. Global burden of disease attributable to mental and substance use disorders: Findings from the global burden of disease study 2010. *The Lancet* 382, 9904 (2010), 1575–1586. DOI: https://doi.org/10.1016/S0140-6736(13)61611-6

[108] Paula Wilbourne, Geralyn Dexter, and David Shoup. 2018. Research driven: Sibly and the transformation of mental health and wellness. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 389–391. DOI: https://doi.org/10.1145/3240925.3240932

[109] Christine T. Wolf. 2019. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 252–257. DOI: https://doi.org/10.1145/3301275.3302317

[110] Jesse H. Wright, Jesse J. Owen, Derek Richards, Tracy D. Eells, Thomas Richardson, Gregory K. Brown, Marna Barrett, Mary Ann Rasku, Geneva Polser, and Michael E. Thase. 2019. Computer-Assisted Cognitive-Behavior Therapy for Depression: A Systematic Review and Meta-Analysis. *The Journal of Clinical Psychiatry* 80, 2 (2019), 18r12188. DOI: https://doi.org/10.4088/JCP.18r12188

[111] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–11. DOI: https://doi.org/10.1145/3290605.3300468

[112] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. Jana Diesner, Elena Ferrari, and Guandong Xu (Eds.), ACM, 1191–1198. DOI: https://doi.org/10.1145/3110025.3123028

[113] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2, 10 (2018), 719–731. DOI: https://doi.org/10.1038/s41551-018-0305-z

[114] Camellia Zakaria, Rajesh Balan, and Youngki Lee. 2019. StressMon: Scalable detection of perceived stress and depression using passive sensing of changes in work routines and group interactions. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 29 pages. DOI: https://doi.org/10.1145/3359139

[115] Sigal Zilcha-Mano. 2017. Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *American Psychologist* 72, 4 (2017), 311–325. DOI: https://doi.org/10.1037/a0040435

Received 12 November 2021; revised 27 May 2022; accepted 25 August 2022