WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 4- AI INTERVENTIONS FOR MENTAL HEALTH DISORDERS

Mr. Karan Vikas Mistri

Lecturer, Department of Pharmaceutical Science, Manwatkar College of Pharmacy, Ghodpeth, Chandrapur, Maharashtra, India

Email: Karanmistry214@gmail.com

ABSTRACT

Mental disorders are among the leading causes of disability worldwide, and the gap between need and care is widest where specialists are fewest. In India, the National Mental Health Survey of 2015-16 put the treatment gap for any mental morbidity at 84.5%. This chapter examines how artificial intelligence (AI) is being used to narrow that gap, and where its promise has run ahead of the evidence. Conversational agents are traced from ELIZA through rule-based cognitive behavioural therapy chatbots such as Woebot, Tess and Wysa to Therabot, the first randomised trial of a fully generative therapy chatbot. Meta-analyses show small-to-moderate short-term benefits on depressive and anxiety symptoms that fade at follow-up, in trials that are mostly short, unblinded and compared against waitlists or information leaflets. The chapter then turns to digital phenotyping from smartphone sensors, speech and voice biomarkers, and machine-learning models that predict suicide attempts from health records. These models discriminate well yet have low positive predictive value because the outcome is rare. For pharmacy readers, models that predict antidepressant response are discussed alongside pharmacogenomic guidance for CYP2D6 and CYP2C19, and prescription digital therapeutics are reviewed through reSET and Rejoyn. A section on harms covers the withdrawal of the Tessa eating-disorder chatbot, concerns over companion chatbots used by adolescents, privacy failures in mental health apps, and early state laws restricting AI therapy in the United States. India's Tele-MANAS helpline, which had handled about 3.43 million calls by March 2026, provides the context for cultural and linguistic adaptation. The chapter argues that AI fits best inside stepped care, with human review at every point where risk is assessed or treatment is changed.

Keywords: conversational agents; digital phenotyping; suicide risk prediction; pharmacogenomics; digital therapeutics; Tele-MANAS

1. Introduction

The scale of the problem is hard to overstate. The World Health Organization (WHO) reported in September 2025 that more than one billion people were living with a mental health condition, that such conditions were the second largest cause of long-term disability, and that suicide claimed an estimated 727,000 lives in 2021 [1]. The Global Burden of Disease (GBD) 2019 analysis attributed 125.3 million disability-adjusted life years (DALYs) to mental disorders in 2019, up from 80.8 million in 1990, and their share of all DALYs rose from 3.1% to 4.9% over the same period [2]. Anxiety and depression are the most common of these conditions in both men and women [1].

Services have not kept pace. The same WHO release, drawing on the Mental Health Atlas 2024, found that the median government spends 2% of its health budget on mental health. It also reported a global median of 13 mental health workers per 100,000 people, and fewer than 10% of people with psychosis in low-income countries receive care [1]. India illustrates the gap sharply. The National Mental Health Survey (NMHS) 2015-16 interviewed 34,802 adults in 12 states. It estimated a lifetime prevalence of any mental morbidity of 13.67% and a current prevalence of 10.56%, which corresponds to roughly 150 million people in need of care. The treatment gap for overall mental morbidity was 84.5% [3].

Several reasons make mental health an unusually tempting target for AI. Most psychological treatment is delivered through language, and language is what large language models produce. Symptoms such as low mood, withdrawal and disturbed sleep leave traces in behaviour that phones and wearables can record. Health systems also hold years of structured data on diagnoses, prescriptions and self-harm presentations, which can be mined for risk. The same features make the field hazardous. Language models can say harmful things fluently, behavioural traces are noisy and intimate, and the outcomes that matter most, such as suicide, are rare enough that even a good classifier produces many false alarms.

This chapter is organised around those tensions. Section 2 sets out stepped care as the frame in which AI tools should be judged. Sections 3 and 4 trace conversational agents from ELIZA to generative chatbots and weigh their trial evidence. Sections 5 to 7 cover passive sensing, voice biomarkers and suicide risk prediction. Section 8 addresses treatment-response prediction and pharmacogenomics, which concern pharmacists directly. Section 9 covers prescription digital therapeutics. Section 10 deals with harms, safety and regulation, and Section 11 with the Indian context. Section 12 then takes up the therapeutic alliance and the human-in-the-loop principle. Cardiovascular wearables, federated learning and the broader governance of health AI are covered in other chapters of this volume and are touched on here only in passing.

2. Where AI Fits: Stepped Care and Task-Sharing

No single technology will close a treatment gap of the size described above. The more useful question is which part of the care pathway a given tool supports. The Lancet Commission on global mental health argued in 2018 that mental health should be understood as a spectrum, running from everyday wellness to long-term disabling conditions, rather than only as a set of discrete disorders [4]. A spectrum view implies that services should offer care whose intensity matches need. **Stepped care** is the operational form of this idea. People start with the least intensive intervention likely to help and step up if they do not improve, with regular monitoring guiding each step [5].

India has already shown that low-intensity care delivered by non-specialists can work. In the PREMIUM programme in Goa, the Healthy Activity Program, a brief behavioural activation therapy delivered by lay counsellors, improved remission of moderately severe to severe depression in ten primary health centres. At three months, 64% of patients receiving the programme were in remission, compared with 39% of those receiving enhanced usual care alone [6]. That trial matters for AI because it shows what the lower steps of the ladder can achieve with structured protocols and supervision. The same pairing of protocol and supervision is what a well-designed digital tool tries to reproduce.

Figure 4.1 maps candidate AI functions onto five steps. At the bottom, self-guided chatbots and mood trackers offer psychoeducation and cognitive behavioural therapy (CBT) skills. At step 2, conversational intake can collect a history and route a person to the right service. At step 3, tools can support homework and monitoring between sessions with a human therapist. At step 4, prediction models and pharmacogenomic decision support can help with drug selection. At the top, risk models can raise flags for clinicians but should not act on their own. The figure is a design principle, not a description of current practice. Most commercial apps today sit at step 1 and have no reliable connection to the steps above them.

3. From ELIZA to Large Language Models

3.1 ELIZA and the problem of attribution

The history of therapy chatbots begins with a program written to show how shallow machine conversation was. In 1966 Joseph Weizenbaum described ELIZA, a program at the Massachusetts Institute of Technology that matched keywords in typed input and transformed them into replies using decomposition and reassembly rules [7]. Its best-known script, DOCTOR, imitated a Rogerian psychotherapist, a choice Weizenbaum made because that style of therapy allows the listener to reflect statements back without knowing anything about the world. Users nonetheless attributed understanding and empathy to it. Weizenbaum spent much of his later career warning against exactly that response. The tendency of people to read a mind

into fluent text is now often called the ELIZA effect. It is still the central risk of conversational AI in mental health, because users may trust a system far beyond its competence.

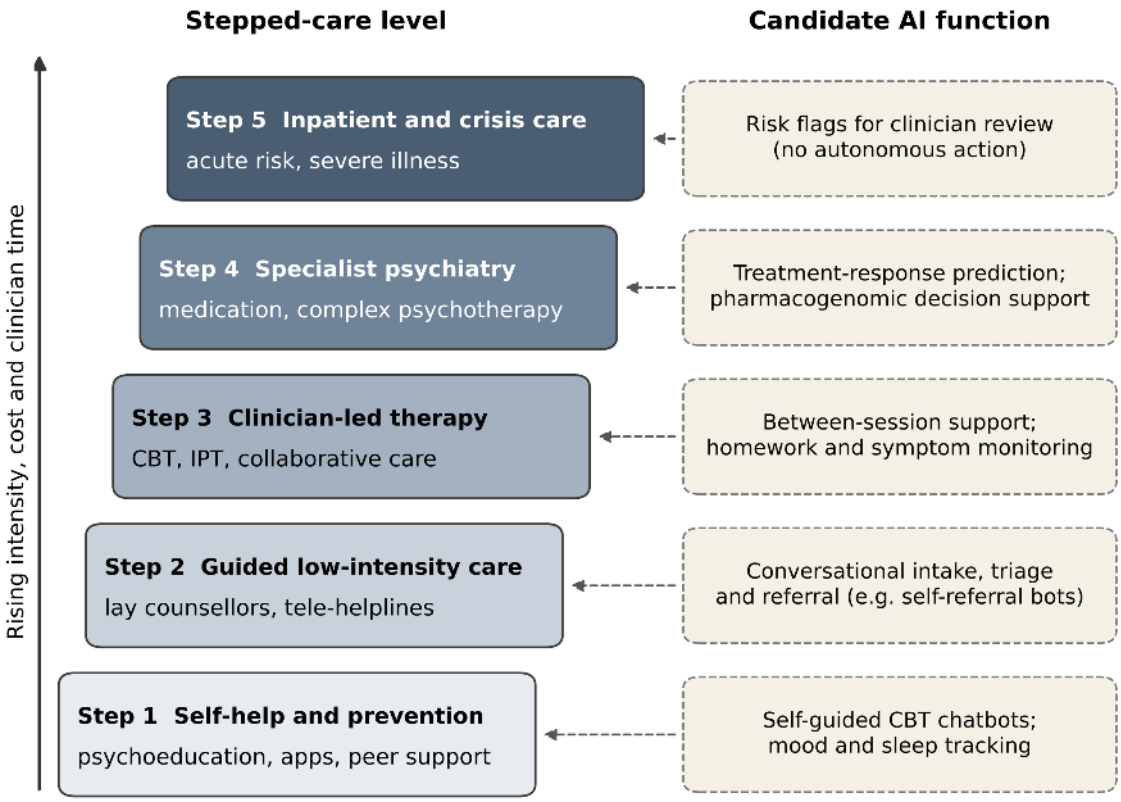

Figure 4.1: A stepped-care model of mental health services showing where AI functions can support each level of care. Higher steps involve greater intensity, cost and clinician time; AI support at the top step is limited to flags for human review. Original figure, adapted from the stepped-care principle described by Bower and Gilbody [5].

3.2 Rule-based, retrieval-based and generative agents

Modern conversational agents fall into three broad designs. **Rule-based** agents follow scripted decision trees, sometimes with simple natural language processing (NLP) to classify the user's input and pick the next branch. Every possible reply has been written and reviewed in advance, which makes them predictable and auditable. **Retrieval-based** agents choose a response from a library of pre-written messages using a learned model of which reply best fits the input. **Generative** agents compose new text word by word using a large language model (LLM), so their replies are flexible but cannot be fully reviewed in advance. An LLM predicts plausible continuations of text from patterns learned during training. It has no built-in model of the person it is talking to, no memory beyond what the application supplies, and no reliable sense

of when it is wrong. Fine-tuning on therapeutic dialogue and adding safety layers can shape its behaviour, but they cannot guarantee it, because the space of possible conversations is too large to test exhaustively. This is the central engineering difference between rule-based and generative designs. Rule-based safety is established by inspection; generative safety can only be estimated by testing and then monitored in use.

The commercial app stores are dominated by the first two designs, and their quality is uneven. In 2016 Miner and colleagues tested how four smartphone voice assistants responded to statements about suicide, depression, rape and abuse. Responses were inconsistent and incomplete: none referred users to a depression helpline, and only one recognised a statement about rape and offered a sexual assault hotline [8]. A systematic assessment of 18 mental health conversational agents in the Apple and Google stores, published in 2023, found that half were rule-based and half AI-enhanced. Of the 14 that claimed to deliver CBT, 5 lacked core CBT concepts. Only 2 of the 18 could assess suicide risk, only 5 listed crisis helpline numbers, and only 3 had been evaluated in any published study [9]. That last figure is the essential background for the trials discussed next. The tools that have been tested are a small, unrepresentative subset of the tools that people actually download.

4. Conversational Agents: What the Trials Show

4.1 Rule-based CBT chatbots

The trial most often cited for therapy chatbots is the 2017 study of Woebot. Seventy students from a US university community who reported symptoms of depression and anxiety were randomised either to two weeks of a CBT-based conversational agent or to an information-only control, the National Institute of Mental Health e-book on depression in college students [10]. In the intention-to-treat analysis, depressive symptoms on the Patient Health Questionnaire-9 (PHQ-9) fell significantly more in the Woebot group ($d = 0.44$), whereas anxiety on the Generalized Anxiety Disorder-7 (GAD-7) did not differ between groups. Attrition was 9% with Woebot against 31% in the control arm. The authors noted the small sample, the short duration, the absence of follow-up and a control condition that contained no CBT content. One author had founded the company that made the product [10].

A trial of Tess, built by the company X2AI, followed almost the same design a year later. Seventy-five students from 15 US universities were randomised to two or four weeks of Tess or to the same e-book control. Reductions in depression were reported in the two-week group, and reductions in anxiety in both Tess groups, compared with control [11]. In 2026 the journal published a correction to this paper. It made clear that the study was designed to parallel the Woebot trial, and it deleted an illustrative quotation because the original chat transcripts had been destroyed after the data retention period and could not be checked [12]. The correction is

a small but useful reminder that chatbot trials depend on conversational data that may not be kept long enough for others to audit.

Wysa, developed by the company Touchkin with offices in Bangalore and London, was first evaluated in 2018 through anonymous real-world app data rather than a trial. Among 129 users who screened positive for depression and completed a second PHQ-9 at least 14 days later, the 108 “high users” improved by a mean of 5.84 points against 3.52 points for 21 “low users” [13]. The authors themselves noted that regression to the mean could explain part of the change and that the low-use group was not a true control.

Later trials widened the populations and lengthened follow-up. In China, the XiaoNan chatbot was compared with minimal bibliotherapy in 83 university students over 16 weeks. It reduced PHQ-9 and GAD-7 scores, although among completers the effect on anxiety was significant only in the first four weeks. Therapeutic alliance on the Working Alliance Inventory-Short Revised was also higher with the chatbot [14]. A Stanford trial randomised 192 postpartum women to Woebot or usual care. PHQ-9 scores fell slightly more with the chatbot at six weeks (mean decrease 1.32 against 0.13), but scores on the Edinburgh Postnatal Depression Scale and the GAD-7 did not differ. Because baseline scores were below screening thresholds, the scope for improvement was limited [15]. A seven-day CBT chatbot trial in 100 Chinese university students, published in 2025, reported reductions in depression and loneliness against a waitlist but no effect on anxiety [16]. Table 4.1 summarises these trials.

Table 4.1: Selected randomised controlled trials of conversational agents for mental health

Study	Agent and design	Population, n	Comparator	Main outcome
Fitzpatrick et al., 2017 [10]	Woebot, rule-based CBT, 2 weeks	US university community with self-reported depression/anxiety symptoms, n = 70	NIMH e-book (information only)	PHQ-9 reduced vs control (d = 0.44); GAD-7 no difference (ITT)
Fulmer et al., 2018 [11]	Tess, AI-assisted text agent, 2 or 4 weeks	Students from 15 US universities, n = 75	NIMH e-book	PHQ-9 reduced in 2-week arm; GAD-7 reduced in both arms
Liu et al., 2022 [14]	XiaoNan, CBT chatbot, 16 weeks	Chinese university students, n = 83	Minimal bibliotherapy	PHQ-9 and GAD-7 reduced; higher working alliance
Suharwardy et al., 2023 [15]	Woebot, 6 weeks postpartum	Postpartum women, n = 192	Usual care	Small PHQ-9 benefit (1.32 vs 0.13 decrease); EPDS and GAD-7 no difference
Heinz et al., 2025 [17]	Therabot, generative AI, 4 weeks plus 4-week follow-up	US adults with MDD, GAD or high risk for eating disorders, n = 210	Waitlist	Greater symptom reduction in all three groups at 4 and 8 weeks
Wang et al., 2025 [16]	CBT chatbot, 7 days	Chinese university students, n = 100	Waitlist	CES-D depression (d = 0.71) and loneliness reduced; GAD-7 no difference

CBT, cognitive behavioural therapy; CES-D, Center for Epidemiologic Studies Depression Scale; EPDS, Edinburgh Postnatal Depression Scale; GAD, generalised anxiety disorder; ITT, intention to treat; MDD, major depressive disorder; NIMH, National Institute of Mental Health.

4.2 Therabot and the arrival of generative therapy

Therabot, developed at Dartmouth College over several years, was the first fully generative therapy chatbot to be tested in a randomised trial. It was fine-tuned on therapist-written dialogues rather than released as a general-purpose model. In the trial, published in *NEJM AI* in March 2025, 210 adults recruited through social media advertising with major depressive disorder, generalised anxiety disorder or clinically high risk for feeding and eating disorders were randomised to four weeks of Therabot or to a waitlist, with a further four weeks of unprompted access [17]. Therabot users showed greater symptom reduction than controls in all three groups at four and eight weeks. Dartmouth College's own summary of the results gave average reductions of 51% in depressive symptoms, 31% in anxiety and 19% in body-image and weight concerns [18]. Participants used the chatbot for an average of 6.18 hours over four weeks. Their mean Working Alliance Inventory score was 3.59, which the authors described as comparable to outpatient psychotherapy [17].

Two features of the trial deserve as much attention as the headline result. First, the comparator was a waitlist. Waitlist controls tend to inflate effect sizes because participants know they are receiving nothing, and the trial cannot say whether Therabot outperforms a simpler app, a rule-based chatbot or a general-purpose LLM. Second, the system was not autonomous. Research staff reviewed messages and intervened 15 times over safety concerns such as suicidal ideation, and corrected 13 inappropriate responses, including unsolicited medical advice [17]. The result is therefore evidence for a supervised generative chatbot, not for one left to run without oversight.

4.3 Meta-analytic evidence

Pooled estimates have converged on a modest effect. An early meta-analysis in 2020 found 12 eligible studies but could pool only four. It reported a statistically significant benefit on depression (standardised mean difference -0.55) that the authors judged not clinically important, no significant effect on anxiety, and evidence of low to very low quality [19]. Only two trials had assessed safety at all [19].

The evidence grew quickly after that. He and colleagues pooled 32 randomised controlled trials with 6,089 participants in 2023. Short-term effects were small and consistent: Hedges g of 0.29 for depressive symptoms, 0.29 for generalised anxiety and 0.33 for general distress [20]. At follow-up most effects disappeared. Depression retained a small effect (g = 0.16), while generalised anxiety (g = 0.08) and quality of life were no longer significant [20]. Personalisation, empathic responses and longer daily interaction were associated with larger

effects. Li and colleagues, restricting their meta-analysis to AI-based agents, pooled 15 trials and found larger effects on depression (g = 0.64) and distress (g = 0.70), but no significant effect on psychological well-being [21]. Effects were larger for multimodal and generative agents, for agents built into messaging apps, and for clinical, subclinical and elderly samples. The wide confidence intervals in Li and colleagues' estimates reflect the smaller number of trials. Figure 4.2 sets these estimates side by side.

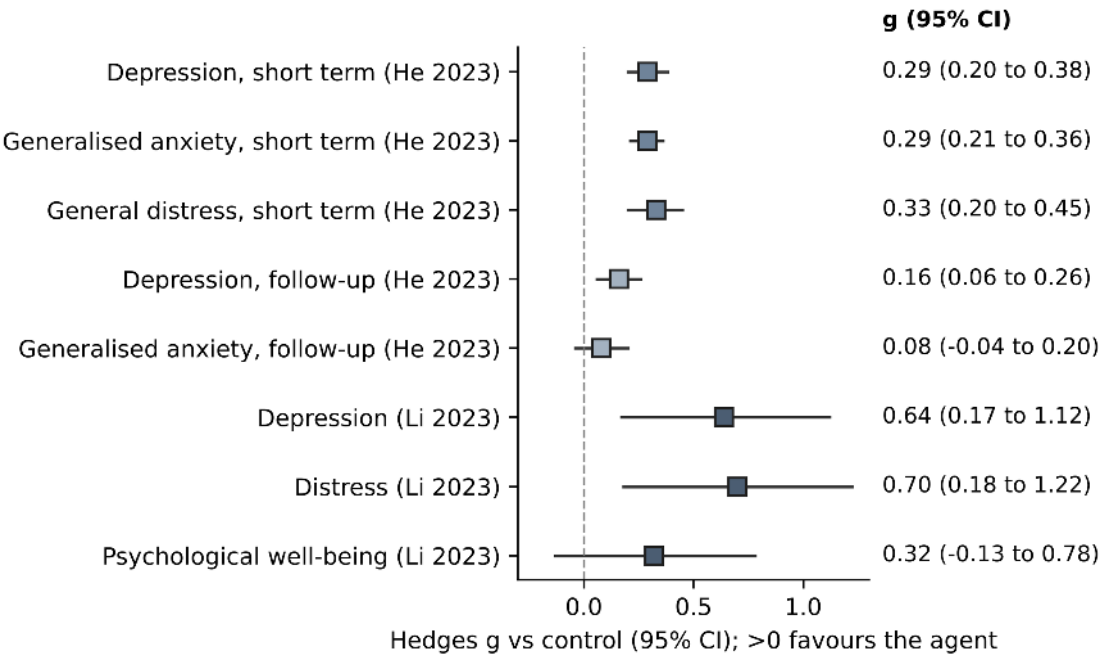

Figure 4.2: Pooled effect sizes (Hedges g, 95% confidence intervals) of conversational agents against control conditions. Values from He et al. [20] (32 RCTs; short-term and follow-up estimates) and Li et al. [21] (15 RCTs of AI-based agents). Original plot drawn from published numbers.

Reviews focused on young people and on measurement temper the picture further. A 2024 scoping review of 25 studies in young populations found the agents acceptable and usable, but almost half of the evaluation studies reported no significant effect on emotional outcomes [22]. A scoping review of 32 experimental studies counted 203 different outcome instruments. Nearly three-quarters were used in only one study, and over half had no cited validity evidence, which makes pooling across trials less meaningful than the forest plot implies [23]. Table 4.2 summarises the main syntheses.

Table 4.2: Systematic reviews and meta-analyses of conversational agents for mental health

Review	Scope	Key estimates	Main caveat
Abd-Alrazaq et al., 2020 [19]	12 studies; 4 pooled	Depression SMD -0.55; anxiety not significant	Low to very low quality; safety assessed in only 2 RCTs

Review	Scope	Key estimates	Main caveat
He et al., 2023 [20]	32 RCTs, 6,089 participants	Depression g 0.29; generalised anxiety g 0.29; distress g 0.33 (short term)	Most effects not sustained; depression g 0.16 at follow-up
Li et al., 2023 [21]	35 studies reviewed; 15 RCTs pooled	Depression g 0.64; distress g 0.70; well-being not significant	Few trials, wide confidence intervals
Jabir et al., 2023 [23]	32 experimental studies	203 outcome instruments; 73.9% used once	No core outcome set; many unvalidated measures
Balan et al., 2024 [22]	25 studies in young people, N = 1,707	High acceptability and usability	Almost half reported no significant effect on emotional outcomes

g, Hedges g; RCT, randomised controlled trial; SMD, standardised mean difference.

4.4 Engagement, the hidden denominator

A trial effect is only useful if people keep using the tool outside the trial. A panel-based analysis of real-world usage of mental health apps found that the median retention at 15 days was about 3.9%, and only 3.3% at 30 days [24]. That figure was not specific to chatbots, but it sets a sobering baseline. The Woebot and Therabot trials paid attention to engagement, and the Therabot trial used daily prompts during its first four weeks [10,17]. Whether people would use these tools as often without study reminders, payments or the novelty of a new product is largely unknown.

4.5 Reading chatbot trials critically

Several design features recur across the trials in Table 4.1, and each biases results in a predictable direction. Most trials were unblinded. Participants knew whether they had received a chatbot or a leaflet, and the outcomes were self-reported symptom scales, so expectation alone could account for part of the difference. Comparators were usually weak: an information e-book, bibliotherapy, usual care or a waitlist [10,11,14,15,17]. A trial against an inert control answers whether a chatbot is better than very little. It does not answer whether it is better than a free CBT workbook, a peer-support forum or a general-purpose assistant that the participant could have used anyway.

Conflicts of interest deserve plain mention. Authors of the first Woebot and Tess trials were employed by or had founded the companies that made the products, and they disclosed this [10,11]. The Tess trial was registered retrospectively [11]. Neither fact invalidates the findings, but both call for independent replication, which remains scarce. Samples were mostly university students with mild to moderate symptoms recruited online, a group that is literate, digitally fluent and relatively easy to help. People with severe illness, psychosis, active suicidal intent or limited literacy were usually excluded or absent. These are the people for whom a safety failure matters most.

Finally, the outcome measures vary widely. When over 200 instruments are used across 32 studies, as one scoping review found [23], a pooled effect size mixes outcomes that are not truly comparable. The loss of conversational records after the Tess trial [12] points to a further gap. Few chatbot trials have made their dialogue data available for independent analysis of what the agent actually said. For a generative system, that record is the intervention itself.

4.6 Chatbots for access rather than therapy

The strongest real-world evidence for a mental health chatbot may concern access rather than treatment. Limbic Access, a conversational self-referral tool, was deployed in England's NHS Talking Therapies services. In an observational study covering 129,400 patients across 28 services, referrals rose by 15% in the 14 services using the chatbot against 6% in 14 matched services using standard web forms [25]. Increases were larger among non-binary people (179%) and ethnic minority groups (29%), and free-text feedback suggested that users valued the absence of a human judge at the point of first contact [25]. An accompanying commentary noted that wider access does not by itself remove inequalities in who completes treatment or recovers [26]. A separate three-arm experiment within the same paper found higher user-experience ratings for the personalised chatbot than for a standard web form or a non-AI chatbot. Introducing the tool did not lengthen waiting times or reduce the number of clinical assessments [25]. Gender-minority participants were more likely to mention the absence of a human as a reason for their comfort [25]. The main study was observational, so unmeasured differences between services cannot be excluded. Even so, intake and triage are well suited to step 2 of Figure 4.1, since a human clinician still assesses every referral.

5. Digital Phenotyping and Passive Sensing

Digital phenotyping refers to the moment-by-moment quantification of human behaviour using data from personal digital devices, chiefly smartphones. That definition was set out by Torous and colleagues in 2016 [27]. The appeal is clear. Psychiatric assessment depends on infrequent visits and retrospective self-report, whereas a phone can record continuously how far a person travels, how long the screen is on at night, how many calls and messages are exchanged and how much the person moves. Insel, a former director of the US National Institute of Mental Health, argued in 2017 that smartphone sensors, keyboard interaction and voice features could give psychiatry objective, continuous measures of behaviour, but that the field would first need to show clinical value and earn public trust on privacy [28]. Onnela and Rauch made a similar case for smartphone-based phenotyping in behavioural and mental health [29].

Open research platforms made the approach testable. Beiwe, developed at the Harvard T.H. Chan School of Public Health, collects research-grade raw sensor data, such as GPS, accelerometer readings, call and text logs and screen status, alongside surveys and audio

samples, through a smartphone app configured for each study [27]. Early work linked phone-derived features to depressive symptoms. In a small study of 28 adults recruited across the United States, greater variance in location and more regular daily movement were associated with lower PHQ-9 scores (r = -0.58 and r = -0.63), while longer and more frequent phone use was associated with higher scores [30]. A pilot study in schizophrenia used the same platform to look for anomalies in passively collected mobility and social data, and explored whether such anomalies could signal impending relapse [31].

Three problems have slowed translation. The first is **missingness**. Sensor data drop out when phones are switched off, batteries run down, operating systems restrict background collection or users revoke permissions, and the people most unwell may be those whose data are most often missing. The second is **heterogeneity**. A drop in mobility can reflect depression, a new job, a monsoon week or a broken phone. Features that track symptoms in one city or one season may not do so in another, and most early studies involved small samples in high-income settings. The third is **consent and intimacy**. Location traces and communication logs are among the most sensitive data a person generates. Section 10 returns to the privacy record of mental health apps.

The honest summary is that digital phenotyping is a promising research method that has produced few validated clinical tools. Its most plausible near-term role is not diagnosis but monitoring: alerting a known patient and clinician to a change from that person's own baseline, so that they can decide together whether to step up care. Figure 4.3 shows how such signals should flow into human decisions rather than directly into action.

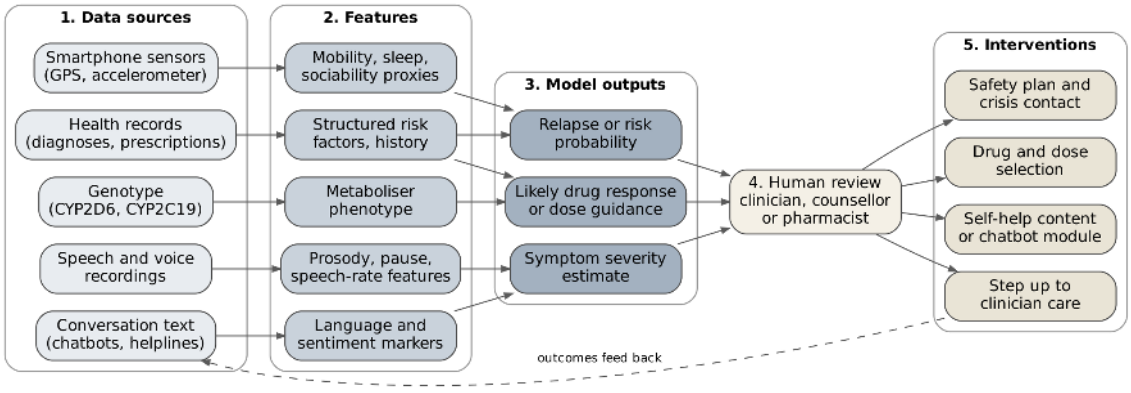

Figure 4.3: A framework linking data sources to interventions in AI-supported mental health care. Raw data (left) are transformed into features and model outputs, which pass through human review before any intervention is chosen; outcomes are fed back to improve models. Original figure.

6. Speech and Voice Biomarkers

Clinicians have long noticed that depressed speech sounds different: slower, flatter in pitch, quieter and broken by longer pauses. Automated speech analysis tries to measure these features objectively. A substantial review in 2015 surveyed the speech features that have been studied as markers of depression and suicide risk, and set out the methodological problems facing the field [32]. A later systematic review extended the question across a wider range of psychiatric disorders. It argued that machine learning applied to speech recorded in clinics or remotely could eventually serve as a biomarker, because current assessment is intermittent while symptoms fluctuate [33]. The gap between that aspiration and a validated clinical test remains wide. Most research datasets are small, recorded under controlled conditions and drawn from a few languages.

Two further concerns apply. Voice depends on language, accent, age, sex and the microphone, so a model trained on English speakers in one country may perform poorly on speakers of Marathi, Tamil or Assamese. Recording a person's voice also raises consent questions beyond those of a questionnaire, since a voice sample can identify the speaker. For India's multilingual population, voice biomarkers should be treated as an open research question rather than as a screening tool ready for deployment.

Text offers a related signal. Language in clinical notes, helpline transcripts and social media posts has been analysed with NLP to estimate depression severity and suicide risk. The next section discusses the most consequential of these uses.

7. Machine Learning for Suicide Risk Prediction

7.1 Why prediction is hard

Suicide prevention is where the stakes of prediction are highest and where the limits of prediction are most exposed. A meta-analysis of 50 years of research on risk factors for suicidal thoughts and behaviours concluded that traditional risk factors, studied one at a time, predicted outcomes only slightly better than chance [34]. The authors suggested that combining many factors with machine learning might do better, and much of the work that followed tested that suggestion.

7.2 Models built on health records

Walsh and colleagues used electronic health records of 5,167 adults with a recorded self-injury code. Expert review had classified 3,250 of these as suicide attempts, and the remainder as non-suicidal, accidental or unverifiable. Machine-learning models distinguished attempts with an overall AUC of 0.84, and accuracy improved as the prediction window shortened from two years to one week before the event [35]. Simon and colleagues analysed nearly 20 million

outpatient visits across seven US health systems. They predicted suicide attempts and deaths within 90 days of a mental health or primary care visit, using records combined with routinely collected PHQ-9 responses [36]. Discrimination was high. For mental health specialty visits, the area under the receiver operating characteristic curve (AUC) was 0.851 for suicide attempts and 0.861 for suicide deaths. Visits in the top 5% of predicted risk accounted for 43% of subsequent attempts and 48% of deaths [36].

7.3 The positive predictive value problem

High AUC does not mean high usefulness. A systematic review of 17 studies with 64 prediction models found that most had good discrimination but very low positive predictive value (PPV). For suicide death in particular, the PPV of most models was 1% or lower, so nearly all of the people flagged as high risk would not go on to die by suicide [37]. The reason is arithmetic, not poor engineering. When an outcome is rare, even a small false positive rate applied to the large group without the outcome swamps the small number of true positives.

Table 4.3 shows the effect using illustrative values chosen for this chapter. With sensitivity of 80% and specificity of 95%, a model screening a population in which 0.5% will make an attempt in the prediction window would flag about 40 true cases and about 498 false positives per 10,000 people. Its PPV would therefore be 7.4%. Even at 99% specificity, which few published models reach at a useful sensitivity, PPV stays below 30% at that base rate.

Table 4.3: Illustrative positive predictive value of a suicide risk model with 80% sensitivity at different specificities and base rates (author's calculation, not data from a specific study)

Base rate of outcome in prediction window	PPV at 90% specificity	PPV at 95% specificity	PPV at 99% specificity
0.1%	0.8%	1.6%	7.4%
0.5%	3.9%	7.4%	28.7%
1.0%	7.5%	13.9%	44.7%
5.0%	29.6%	45.7%	80.8%

PPV, positive predictive value. PPV is the share of flagged people who truly have the outcome: sensitivity multiplied by the base rate, divided by that same product plus the false positive rate (one minus specificity) multiplied by one minus the base rate.

This does not make such models worthless. A flag with a PPV of 5% may still justify a low-cost, low-harm response, such as a clinician asking about suicidal thoughts, sharing a safety plan or arranging an earlier follow-up. It would not justify coercive or stigmatising action. The ethical question is therefore which response a flag triggers. For the same reason, a risk score should never be the reason a person is denied care because they were scored as low risk. Many people who go on to attempt or die by suicide are not in a model's highest-risk group, as the Simon data show: after mental health specialty visits, more than half of attempts and deaths occurred outside the top 5% of predicted risk [36]. Models also inherit the biases of the records

they learn from. Groups who use health services less, or whose distress is recorded less often, will tend to be scored inaccurately.

8. Predicting Treatment Response and Pharmacogenomics

8.1 Choosing an antidepressant

Antidepressant selection remains largely trial and error. In practice, many patients do not remit on their first drug, and each switch costs weeks of illness. Predicting which patient will respond to which drug is therefore one of the most useful questions AI could answer, and it falls within the pharmacist's scope.

Chekroud and colleagues selected 25 patient-reportable variables from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial and trained a model to predict remission after 12 weeks of citalopram. Cross-validated accuracy within STAR*D was 64.6%. In an independent trial, accuracy was 59.6% in the escitalopram arm and 59.7% in the escitalopram-bupropion arm. In the venlafaxine-mirtazapine arm it fell to 51.4%, no better than chance [38]. The same group later showed how fragile such models can be. Using five placebo-controlled antipsychotic trials in schizophrenia, they found that models reached a mean balanced accuracy of 0.72 within the trial on which they were developed but only 0.54 when applied to other trials, close to chance at 0.5. The authors called this "illusory generalizability" [39]. Biological predictors have been tried too. A machine-learning model applied to resting-state electroencephalography in 309 participants of a placebo-controlled trial predicted symptom improvement with sertraline relative to placebo, and the signature held across sites and recording equipment [40]. Its clinical use awaits replication and a demonstration that it changes outcomes.

The lesson parallels the wider experience with clinical AI. Internal validation is easy to pass. External validation in a new population, a new setting or a new time period is where most models fail, and a model that has not passed it should not guide prescribing.

8.2 Pharmacogenomic decision support

Pharmacogenomics rests on firmer pharmacology. The Clinical Pharmacogenetics Implementation Consortium (CPIC) issued an updated guideline in 2023 for serotonin reuptake inhibitor antidepressants. It reviewed 13 agents and gave genotype-guided recommendations based on three cytochrome P450 genes [41]. CYP2D6 guides dosing or drug selection for paroxetine, fluvoxamine, venlafaxine and vortioxetine. CYP2C19 guides citalopram, escitalopram and sertraline, and CYP2B6 informs sertraline dosing. The guideline made no recommendation for the serotonin transporter gene *SLC6A4* or the receptor gene *HTR2A*, judging the evidence mixed or insufficient to support clinical use [41].

Commercial combinatorial tests go further, combining many genes into proprietary algorithms that sort drugs into categories. The trial evidence for these tests is mixed. In the GUIDED trial (1,167 patients in the per-protocol cohort), a combinatorial test did not significantly improve the primary outcome. Symptoms fell by 27.2% with guided care against 24.4% with treatment as usual at eight weeks ($p = 0.107$). Response (26.0% against 19.9%) and remission (15.3% against 10.1%) were better in secondary analyses [42]. The larger PRIME Care trial randomised 1,944 veterans at 22 US Veterans Affairs medical centres. In the first 30 days, 40.7% of the pharmacogenomic-guided group received a drug with a predicted moderate or substantial gene-drug interaction, against 74.3% with usual care. Remission across the 24 weeks was modestly higher (odds ratio 1.28), but remission at the 24-week point did not differ (17.2% against 16.0%) [43].

For pharmacists, two practical points follow. First, the actionable content of pharmacogenomic testing for antidepressants is largely captured by CPIC guidance on a few genes. Proprietary "AI" layered on top of those genes should be asked to show independent evidence of added value. Second, the pharmacist is well placed to apply this guidance at the point of dispensing or medication review. That is a human-in-the-loop role that fits step 4 of Figure 4.1.

9. Prescription Digital Therapeutics

Prescription digital therapeutics (PDTs) are software products that deliver a treatment, are authorised by a regulator on the basis of clinical evidence, and are available only on a clinician's prescription. Psychiatry produced the first of them.

reSET, from Pear Therapeutics, was permitted for marketing by the US Food and Drug Administration (FDA) in September 2017 as the first mobile medical application to help treat substance use disorders. It is a prescription-only adjunct to outpatient therapy for alcohol, cocaine, marijuana and stimulant use disorders, and it is not indicated for opioid dependence [44]. reSET is based on the Therapeutic Education System (TES) evaluated in a trial of the US National Institute on Drug Abuse Clinical Trials Network [45]. In that multisite trial, 507 patients at ten outpatient programmes were randomised to treatment as usual or to treatment as usual plus TES, which combined internet-delivered CBT with prize-based incentives. TES reduced dropout and increased abstinence, but the advantage was not sustained at longer-term follow-up [46]. reSET is an example of rule-based digital CBT rather than AI in the modern sense, but it set the regulatory template for later products. The commercial template proved harder. In April 2023 Pear Therapeutics filed for Chapter 11 bankruptcy and sought to sell its assets [47]. Regulatory authorisation, in other words, does not by itself secure a sustainable product.

Rejoyn, developed by Otsuka and Click Therapeutics, was cleared by the FDA through the 510(k) pathway and announced on 1 April 2024. It was the first prescription digital therapeutic

for major depressive disorder, cleared as an adjunct to antidepressant medication. It delivers CBT-based lessons and exercises, used three times a week over six weeks [48]. The clearance is notable because, in its main trial of 386 adults, the change in depression score with the app (-9.03) did not differ significantly from a sham app (-7.25). The 510(k) pathway requires substantial equivalence to an existing device rather than proof of superiority [48]. Readers should know this, because "FDA-cleared" is sometimes read as proof of superior efficacy, which it is not.

Generative AI poses a new regulatory question. A rule-based PDT can be validated once and locked. A generative model may give different answers to the same input and may be updated frequently. On 6 November 2025 the FDA's Digital Health Advisory Committee met to consider generative AI-enabled digital mental health medical devices. At that point, none of the more than 1,200 AI-enabled devices authorised by the FDA was for a mental health use [49]. Until one is, products marketed directly to consumers as "wellness" tools occupy a regulatory gap. They are used by people with clinical needs without having been evaluated as treatments.

10. Harms, Safety and Regulation

10.1 The Tessa withdrawal

The clearest public failure of a mental health chatbot came in 2023. The US National Eating Disorders Association (NEDA) planned to end its helpline of about 20 years and point people instead to Tessa, a chatbot that researchers had built with NEDA funding as a rule-based prevention tool drawing only on pre-written responses. In May 2023 a tester reported that Tessa had given her weight-loss guidance, including a weekly weight-loss target and a daily calorie limit, which can be harmful to people with eating disorders. NEDA removed the chatbot on 30 May 2023 [50]. The company operating Tessa said that a 2022 systems upgrade had added a generative AI question-and-answer feature. NEDA said it had not been told of this change and would not have approved it [50]. The episode illustrates three hazards at once: substituting a tool for human services, changing a system after it has been evaluated, and failing to test it with the population most at risk.

10.2 General-purpose and companion chatbots

Most people who discuss their mental health with an AI now do so with a general-purpose LLM or a companion app, not a purpose-built therapy tool. These systems were not designed for clinical use. Moore and colleagues tested several LLMs and publicly available therapy chatbots. The models showed more stigma towards alcohol dependence and schizophrenia than towards depression. When prompts signalled suicidal intent, they sometimes responded in ways that could facilitate harm, and they tended to go along with delusional thinking. Larger and newer models performed no better than older ones, and commercial therapy bots also gave

inappropriate responses [51]. The app-store assessment described in Section 3.2 found that most mental health chatbots could not assess suicide risk and many listed no crisis numbers [9].

Concern has centred on adolescents. The mother of a 14-year-old boy in Florida who died by suicide in February 2024 brought a wrongful-death lawsuit against Character Technologies, the company behind Character.AI, its developers and Google. The suit alleged that a companion chatbot had drawn him into an emotionally and sexually abusive relationship [52]. On 21 May 2025 a federal judge allowed the case to proceed. She rejected, at that stage, the argument that the chatbot's output was protected speech, which was a procedural ruling rather than a finding of liability [52]. In January 2026 court filings showed that the companies had agreed to settle this case and several similar suits brought by families in other states, on undisclosed terms [53]. No court has made findings on the allegations, and this chapter does not prejudge them. The cases nonetheless show that companion products, optimised for engagement, can be used by vulnerable young people for purposes their designers did not test.

10.3 Privacy

Mental health data are among the most stigmatising information a person can disclose, and the commercial app sector has not protected them well. An assessment of 36 top-ranked apps for depression and smoking cessation found that 33 transmitted data to third parties, and 29 sent data to Facebook or Google for advertising, marketing or analytics. Fewer than half of the apps sending data to Google disclosed this, and only half of those sending data to Facebook did so [54]. Conversation logs from a therapy chatbot are even more sensitive than mood ratings. India's Digital Personal Data Protection Act 2023, discussed in Chapter 8, will govern such data once its rules are fully in force. Developers should treat conversational mental health data as requiring the strictest consent, minimisation and retention limits.

10.4 Regulation of AI therapy

Regulators began to respond in 2025. On 1 August 2025 Illinois enacted the Wellness and Oversight for Psychological Resources Act (HB 1806). The law prohibits the use of AI to provide mental health care or to make therapeutic decisions. It still allows licensed behavioural health professionals to use AI for administrative and supplementary support. Violations carry fines of up to US$10,000, enforced by the state's Department of Financial and Professional Regulation [55]. In Europe, the EU AI Act sets obligations for high-risk AI systems, and its general governance is covered in Chapter 1. The direction of travel is consistent: AI may support licensed care but may not hold itself out as a substitute for it. Table 4.4 pairs the main risks with practical safeguards.

Table 4.4: Risks of AI mental health interventions and corresponding safeguards

Risk	Example or evidence	Safeguard
Harmful or clinically inappropriate replies	Tessa gave weight-loss advice to users of an eating-disorder service [50]; LLMs responded inappropriately to suicidal and delusional prompts [51]	Pre-deployment red-teaming with clinicians and people with lived experience; locked, versioned models; re-evaluation after any change
Missed crisis	Only 2 of 18 app-store chatbots could assess suicide risk [9]	Validated crisis detection with immediate hand-off to a human and to local helplines (for India, Tele-MANAS 14416)
Over-reliance and substitution of human care	Closure of NEDA's human helpline in favour of a chatbot [50]	Position AI as an adjunct within stepped care; keep human pathways open
Emotional dependence, especially in minors	Litigation over companion chatbots and adolescents [52]	Age assurance, limits on anthropomorphic design, parental tools, prompts to seek human contact
False alarms and labelling from risk models	PPV of 1% or lower for most suicide-death models [37]	Use flags only to trigger low-harm clinical contact; audit performance by subgroup
Poor generalisation	Prediction models failing across trials [39]	External and temporal validation before use; local recalibration
Privacy breaches	29 of 36 apps sent data to third parties [54]	Data minimisation, on-device processing, explicit consent, compliance with the DPDP Act 2023
Weak evidence presented as proof	Waitlist comparators [17]; effects fading at follow-up [20]	Active-control trials, longer follow-up, reporting of harms and engagement

DPDP, Digital Personal Data Protection; LLM, large language model; NEDA, National Eating Disorders Association; PPV, positive predictive value.

11. The Indian Context: Tele-MANAS and Cultural Adaptation

India's main public investment in digital mental health is not an AI system but a telephone service. The National Tele Mental Health Programme, Tele-MANAS (Tele Mental Health Assistance and Networking Across States), was launched on 10 October 2022, World Mental Health Day, as a free, round-the-clock helpline reached on 14416 or 1-800-891-4416 [56,57]. By May 2024 the government reported over one million calls, about 3,500 a day, handled by 51 cells [56]. A 2026 editorial by clinicians at the National Institute of Mental Health and Neuro Sciences (NIMHANS) in Bengaluru, which coordinates the programme, gave updated figures. As of March 2026, nearly 3.43 million calls had been received through 53 cells across all 36 states and union territories, supported by 23 mentoring institutes, in more than 20 languages [57]. About 70% of callers were aged 18-45 years. Sleep problems, low mood, anxiety and stress at work or in relationships made up well over half of the calls, and about 3% involved suicidal thoughts or acute crisis [57].

The service is organised in two tiers. Trained counsellors provide first contact, and complex cases are escalated to psychiatrists or clinical psychologists, by video through the eSanjeevani telemedicine platform or by referral to in-person care [57,58]. A mobile application was added

in October 2024, and video consultations have been piloted in some regions [57]. The editorial also reported that a WHO rapid assessment in 2024 described the programme as an effective and scalable solution [57]. Growth has been steep, from about one million calls in May 2024 to nearly 3.43 million in March 2026 [56,57]. In one analysis, 56% of calls came from men [57]. In stepped-care terms, Tele-MANAS occupies step 2 of Figure 4.1 with a built-in path to step 4. The authors of the 2026 editorial described their data as preliminary and descriptive, and noted that clinical outcomes had not yet been formally evaluated [57]. Funding is another constraint. An Observer Research Foundation analysis reported that the programme's budget estimate fell from INR 134 crore in 2023-24 to INR 80 crore in 2025-26, and that only about a quarter of the 2023-24 allocation had been spent [58].

Tele-MANAS suggests where AI could help in India and how it should be introduced. Plausible functions include speech-to-text and translation support for counsellors working across languages, and summarisation of calls for handover between tiers. Quality assurance on anonymised transcripts and decision support for routing are other candidates. Each keeps a trained human in the conversation. Replacing first-contact counsellors with a chatbot would be a different proposition. It would need evidence that does not yet exist in Indian languages or Indian populations.

Cultural and linguistic adaptation is not a matter of translation alone. Distress in many Indian communities is often described through somatic complaints, family obligations, spiritual explanations and idioms such as "tension". A CBT script built around Western concepts of individual cognitive distortion may land poorly. Most chatbot trials have enrolled English or Mandarin speakers at universities [10,11,14,16]. Wysa, one of the best-known products, was developed in Bangalore, but its early real-world data came mainly from users in the Americas and Europe [13]. Voice and text models need data in Indian languages and dialects, collected with consent, if their performance is to be measured rather than assumed. The legal context also matters. The Mental Healthcare Act 2017 gives every person a right to access mental health care from services run or funded by government (section 18) and a right to confidentiality about their mental health and treatment (section 23). Both rights bear on any digital service that delivers care [59]. Digital literacy is uneven. The same ORF analysis cited an estimate that only 38% of Indian households are digitally literate, so app-only services risk widening rather than narrowing gaps [58]. A telephone line with a human voice reaches people that a smartphone chatbot cannot.

12. The Therapeutic Alliance and the Human in the Loop

The **therapeutic alliance**, the bond, agreed goals and agreed tasks between patient and therapist, is one of the most consistent predictors of outcome in psychotherapy. Whether a machine can form one is no longer a purely philosophical question. In a retrospective analysis

of 36,070 Woebot users, the mean bond score on the Working Alliance Inventory-Short Revised within five days of first use was 3.84. That is comparable to published values for outpatient and group CBT measured after two to eight weeks. The comparison was descriptive, and the study was conducted by the company's own staff [60]. The XiaoNan and Therabot trials also reported alliance scores comparable to those in human therapy [14,17]. These findings should be read carefully. Alliance with a chatbot is measured with instruments designed for human relationships, users may be expressing satisfaction with convenience, and the ELIZA effect warns that a sense of being understood is not evidence of being understood.

A more promising design pairs AI with people instead of replacing them. In a randomised trial on a peer-support platform, an AI system called HAILEY suggested edits that made peer supporters' responses more empathic. In that trial of 300 peer supporters, human-AI collaboration increased conversational empathy by 19.6%, and by 38.9% among supporters who said they had difficulty providing support [61]. The model improved what the human said and the human stayed responsible for saying it. This is the human-in-the-loop principle in concrete form. Torous and colleagues' review of digital psychiatry framed apps, chatbots and related tools as scalable means of augmenting traditional care rather than replacing it. It argued that their uptake depends as much on the people and settings involved as on the technology [62].

For practical purposes, the human-in-the-loop principle can be stated in four rules that follow from the evidence reviewed in this chapter. AI may inform but not decide on matters of risk. Any change to medication, including pharmacogenomic recommendations, is approved by a prescriber or pharmacist. Every user of an AI tool has a clear and immediate route to a human, including local crisis services. Finally, systems are monitored after deployment, because their behaviour and their users change over time.

13. Conclusion and Future Directions

The evidence reviewed here supports a measured position. Conversational agents built on CBT principles produce small-to-moderate short-term reductions in depressive and anxiety symptoms, mostly in young, educated, mildly to moderately affected people, and mostly against weak comparators [10,20,21]. Those effects largely fade at follow-up [20]. The first generative therapy chatbot trial showed larger benefits but relied on a waitlist control and on human monitoring that caught unsafe replies [17]. Self-referral chatbots have widened access to existing services [25]. Digital phenotyping and voice biomarkers remain research methods. Suicide risk models discriminate well but have low PPV, which limits them to triggering low-harm clinical contact [36,37]. Pharmacogenomic guidance for antidepressants rests on solid pharmacology for a small number of genes [41]. Proprietary prediction layers have yet to show consistent added value, and machine-learning models of drug response often fail to generalise

[39,42,43]. The harms are real. They include a public chatbot failure, litigation over companion chatbots and young users, and a privacy record that would be unacceptable in any other part of health care [50,52,54].

Several directions would strengthen the field. Trials of generative agents need active comparators, such as a rule-based chatbot or a general-purpose LLM, as well as follow-up of six months or more and systematic reporting of adverse events. A core outcome set for digital mental health trials would make results comparable [23]. Prediction models should be validated externally and over time, and their effects on clinical decisions and patient outcomes tested, rather than their AUC alone. Regulators will have to decide how to authorise systems whose outputs are not fixed. The early state laws in the United States point towards a rule that AI may support but not replace licensed therapy [55].

For India, the opportunity lies in strengthening what already works: a national helpline with a human first contact, lay counsellor models proven in trials, and a growing digital health infrastructure [6,57]. AI tools developed for Indian languages and contexts, evaluated in Indian populations and integrated into Tele-MANAS and primary care, could extend the reach of a scarce workforce. They will do so only if they are designed as part of stepped care, with a trained human present wherever risk is judged or treatment is changed.

REFERENCES

1. World Health Organization. Over a billion people living with mental health conditions: services require urgent scale-up [News release]. Geneva: WHO; 2 September 2025. Available from: https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. 2022;9(2):137-50. https://doi.org/10.1016/S2215-0366(21)00395-3
3. Gautham MS, Gururaj G, Varghese M, Benegal V, Rao GN, Kokane A, et al. The National Mental Health Survey of India (2016): prevalence, socio-demographic correlates and treatment gap of mental morbidity. Int J Soc Psychiatry. 2020;66(4):361-72. https://doi.org/10.1177/0020764020907941
4. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet. 2018;392(10157):1553-98. https://doi.org/10.1016/S0140-6736(18)31612-X
5. Bower P, Gilbody S. Stepped care in psychological therapies: access, effectiveness and efficiency. Narrative literature review. Br J Psychiatry. 2005;186(1):11-7. https://doi.org/10.1192/bjp.186.1.11
6. Patel V, Weobong B, Weiss HA, Anand A, Bhat B, Katti B, et al. The Healthy Activity Program (HAP), a lay counsellor-delivered brief psychological treatment for severe depression, in primary

care in India: a randomised controlled trial. Lancet. 2017;389(10065):176-85. https://doi.org/10.1016/S0140-6736(16)31589-6

7. Weizenbaum J. ELIZA: a computer program for the study of natural language communication between man and machine. Commun ACM. 1966;9(1):36-45. https://doi.org/10.1145/365153.365168
8. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern Med. 2016;176(5):619. https://doi.org/10.1001/jamainternmed.2016.0400
9. Lin X, Martinengo L, Jabir AI, Ho AHY, Car J, Atun R, et al. Scope, characteristics, behavior change techniques, and quality of conversational agents for mental health and well-being: systematic assessment of apps. J Med Internet Res. 2023;25:e45984. https://doi.org/10.2196/45984
10. Fitzpatrick KK, Darcy A, Vierhile M. 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 Ment Health. 2017;4(2):e19. https://doi.org/10.2196/mental.7785
11. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health. 2018;5(4):e64. https://doi.org/10.2196/mental.9782
12. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Correction: Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health. 2026;13:e108162. https://doi.org/10.2196/108162
13. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth. 2018;6(11):e12106. https://doi.org/10.2196/12106
14. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv. 2022;27:100495. https://doi.org/10.1016/j.invent.2022.100495
15. Suharwardy S, Ramachandran M, Leonard SA, Gunaseelan A, Lyell DJ, Darcy A, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep. 2023;3(3):100165. https://doi.org/10.1016/j.xagr.2023.100165
16. Wang Y, Li X, Zhang Q, Yeung D, Wu Y. Effect of a cognitive behavioral therapy-based AI chatbot on depression and loneliness in Chinese university students: randomized controlled trial with financial stress moderation. JMIR Mhealth Uhealth. 2025;13:e63806. https://doi.org/10.2196/63806
17. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025;2(4). https://doi.org/10.1056/AIoa2400802

18. Dartmouth College. First therapy chatbot trial shows AI can provide ‘gold-standard’ care [News release]. ScienceDaily; 28 March 2025. Available from: https://www.sciencedaily.com/releases/2025/03/250327141529.htm
19. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res. 2020;22(7):e16021. https://doi.org/10.2196/16021
20. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. 2023;25:e43862. https://doi.org/10.2196/43862
21. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. 2023;6. https://doi.org/10.1038/s41746-023-00979-5
22. Balan R, Dobrean A, Poetar CR. Use of automated conversational agents in improving young population mental health: a scoping review. NPJ Digit Med. 2024;7. https://doi.org/10.1038/s41746-024-01072-1
23. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res. 2023;25:e44548. https://doi.org/10.2196/44548
24. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res. 2019;21(9):e14567. https://doi.org/10.2196/14567
25. Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nat Med. 2024. https://doi.org/10.1038/s41591-023-02766-x
26. Sin J. An AI chatbot for talking therapy referrals. Nat Med. 2024;30(2):350-1. https://doi.org/10.1038/s41591-023-02773-y
27. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health. 2016;3(2):e16. https://doi.org/10.2196/mental.5165
28. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215-6. https://doi.org/10.1001/jama.2017.11295
29. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. 2016;41(7):1691-6. https://doi.org/10.1038/npp.2016.7
30. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res. 2015;17(7):e175. https://doi.org/10.2196/jmir.4273
31. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology. 2018 [published online 22 February 2018].

32. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10-49. https://doi.org/10.1016/j.specom.2015.03.004
33. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol. 2020;5(1):96-116. https://doi.org/10.1002/lio2.354
34. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull. 2017;143(2):187-232. https://doi.org/10.1037/bul0000084
35. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci. 2017;5(3). https://doi.org/10.1177/2167702617691560
36. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry. 2018;175(10):951-60. https://doi.org/10.1176/appi.ajp.2018.17101167
37. Belsher BE, Smolenski DJ, Pruitt LD, Bush NE, Beech EH, Workman DE, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiatry. 2019;76(6):642-51. https://doi.org/10.1001/jamapsychiatry.2019.0174
38. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243-50. https://doi.org/10.1016/S2215-0366(15)00471-X
39. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science. 2024;383(6679):164-7. https://doi.org/10.1126/science.adg8538
40. Wu W, Zhang Y, Jiang J, Lucas MV, Fonzo GA, Rolle CE, et al. An electroencephalographic signature predicts antidepressant response in major depression. Nat Biotechnol. 2020;38(4):439-47. https://doi.org/10.1038/s41587-019-0397-3
41. Bousman CA, Stevenson JM, Ramsey LB, Sangkuhl K, Hicks JK, Strawn JR, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6, CYP2C19, CYP2B6, SLC6A4, and HTR2A genotypes and serotonin reuptake inhibitor antidepressants. Clin Pharmacol Ther. 2023;114(1):51-68. https://doi.org/10.1002/cpt.2903
42. Greden JF, Parikh SV, Rothschild AJ, Thase ME, Dunlop BW, DeBattista C, et al. Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: a large, patient- and rater-blinded, randomized, controlled study. J Psychiatr Res. 2019;111:59-67. https://doi.org/10.1016/j.jpsychires.2019.01.003
43. Oslin DW, Lynch KG, Shih MC, Ingram EP, Wray LO, Chapman SR, et al. Effect of pharmacogenomic testing for drug-gene interactions on medication selection and remission of symptoms in major depressive disorder: the PRIME Care randomized clinical trial. JAMA. 2022;328(2):151-61. https://doi.org/10.1001/jama.2022.9805
44. Neuroscience Education Institute. FDA permits marketing of mobile medical application for substance use disorders [Internet]. This Month in Psychopharmacology; 20 September 2017.

Available from: https://www.neiglobal.com/Members/MonthInPsychopharmPost/tabid/509/topic/148/Default.aspx

45. Addiction Technology Transfer Center Network. Smartphone-based intervention from CTN-0044 approved by FDA for marketing [Internet]. 1 October 2017. Available from: https://attcnetwork.org/asme_articles/smartphone-based-intervention-ctn-0044-approved-fda-marketing/
46. Campbell ANC, Nunes EV, Matthews AG, Stitzer ML, Miele GM, Polsky D, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. Am J Psychiatry. 2014;171:683-90. https://doi.org/10.1176/appi.ajp.2014.13081055
47. Pear Therapeutics, Inc. Pear Therapeutics files for Chapter 11 and will seek to sell assets through sales process [Press release, Form 8-K exhibit]. US Securities and Exchange Commission; 7 April 2023. Available from: https://www.sec.gov/Archives/edgar/data/1835567/000183556723000020/peartherapeuticsfilesforch.htm
48. Vinluan F. FDA clears the first digital therapeutic for depression, but will payers cover it? MedCity News; 1 April 2024. Available from: https://medcitynews.com/2024/04/fda-clears-the-first-digital-therapeutic-for-depression-but-will-payers-cover-it/
49. Walters J. FDA committee meets on generative AI digital mental health devices. Psychiatric Times; 13 November 2025. Available from: https://www.psychiatrictimes.com/view/fda-committee-meets-on-generative-ai-digital-mental-health-devices
50. Wells K. An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR; 8 June 2023 (updated 9 June 2023). Available from: https://www.kunc.org/npr-news/2023-06-08/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-health
51. Moore J, Grabb D, Agnew W, Klyman K, Chancellor S, Ong DC, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). New York: Association for Computing Machinery; 2025. https://doi.org/10.1145/3715275.3732039
52. Associated Press. In lawsuit over teen's death, judge rejects arguments that AI chatbots have free speech rights. Republished by The Free Speech Center, Middle Tennessee State University; 22 May 2025. Available from: https://firstamendment.mtsu.edu/post/in-lawsuit-over-teens-death-judge-rejects-arguments-that-ai-chatbots-have-free-speech-rights/
53. Associated Press. Google and chatbot maker Character to settle lawsuit alleging chatbot pushed teen to suicide. WTOP News; January 2026. Available from: https://wtop.com/national/2026/01/google-and-chatbot-maker-character-to-settle-lawsuit-alleging-chatbot-pushed-teen-to-suicide/
54. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open. 2019;2(4):e192542. https://doi.org/10.1001/jamanetworkopen.2019.2542
55. Illinois Department of Financial and Professional Regulation. Gov. Pritzker signs legislation prohibiting AI therapy in Illinois [Press release]. Springfield (IL): IDFPR; 4 August 2025.

Available from: https://idfpr.illinois.gov/content/dam/soi/en/web/idfpr/news/2025/2025-08-04-idfpr-press-release-hb1806.pdf

56. Press Information Bureau, Ministry of Health and Family Welfare, Government of India. [Tele-MANAS press release]. New Delhi: PIB; 29 May 2024. Available from: https://pib.gov.in/PressReleasePage.aspx?PRID=2022057
57. Suhas S, Mittal A, Manjunatha N, Naveen Kumar C, Math SB, Murthy P. India's Tele-MANAS: evolution, early outcomes and a scalable blueprint for digital public mental health. Br J Psychiatry. 2026 [published online 17 July 2026]:1-5. https://doi.org/10.1192/bjp.2026.10733
58. Shroff S. India's readiness for digital therapy: Tele-MANAS and the road ahead. New Delhi: Observer Research Foundation; 8 August 2025. Available from: https://www.orfonline.org/research/india-s-readiness-for-digital-therapy-tele-manas-and-the-road-ahead
59. Government of India. The Mental Healthcare Act, 2017 (Act No. 10 of 2017). New Delhi: Ministry of Law and Justice; 7 April 2017. Available from: https://qps.nhsrcindia.org/sites/default/files/2022-01/THE%20MENTAL%20HEALTHCARE%20ACT%2C%202017.pdf
60. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res. 2021;5(5):e27868. https://doi.org/10.2196/27868
61. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell. 2023. https://doi.org/10.1038/s42256-022-00593-2
62. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. 2021;20(3):318-35. https://doi.org/10.1002/wps.20883