PSILOGOS

Research Article

open-access

Auditability and Care Proximity in Artificial Intelligence for Psychological Services: A Study-Level Robustness Analysis

Jongsuk Ro

ABSTRACT

Artificial intelligence has entered psychological services through diagnostic classifiers, clinical-language models, passive sensing, conversational treatment, virtual-reality practice, and remote psychotherapy. Evidence for these uses is commonly discussed by technical family, although a clinical decision also depends on how directly a system reaches a patient, whether the study design permits causal interpretation, whether the endpoint concerns symptoms rather than an intermediate signal, and whether the population size can be audited. This study asks which of those properties governs the robustness of clinical evidence proximity across early applications in psychological care. The analytic cohort comprised nine investigations published between 2011 and 2020. Seven provided exact participant counts, representing 20,116 people; one of these also analysed 90,934 psychotherapy transcripts. Each record was coded for population scale, design strength, care proximity, and endpoint proximity. A weighted clinical evidence proximity score was examined with 100,000 Dirichlet weight perturbations, probabilistic treatment of two unreported participant totals, complete criterion omission, and an exact 126-allocation permutation test. Median scores ranged from 42.4 to 85.4. Automated cognitive behavioural therapy for insomnia and telephone psychotherapy occupied first place in 43.1% and 48.9% of perturbations, respectively, whereas conversational cognitive behavioural therapy entered the first three positions in 99.95%. Patient-facing interventions exceeded the remaining applications by 19.2 points on a care-excluded three-criterion score (exact $p = 0.0238$). Large population size alone did not secure a leading position: the 17,572-patient psychotherapy-language study had a median score of 64.9 because it measured treatment process rather than symptom change. The research question is therefore answered by a joint condition. Direct care delivery was associated with higher evidentiary proximity, but stable leadership required both a comparative design and an endpoint connected to clinical change. The findings support transparent, criterion-specific assessment when psychological artificial intelligence is considered for clinical use.

Keywords: clinical artificial intelligence; psychological intervention; diagnostic classification; digital mental health; evidence audit; sensitivity analysis

Department of Psychiatry, College of Medicine, Seoul National University, Seoul, South Korea
jongsuk.ro81@gmail.com

– **Published:** 11 October 2024

1. INTRODUCTION

Psychological care is being reshaped by computational systems that detect regularities in clinical records, interpret neuroimaging, infer behavioural states from phones and wearables, analyse psychotherapy language, and deliver structured treatment through digital interfaces.[5] This activity responds to a substantial service problem: the demand for assessment and treatment exceeds the availability of trained clinicians in many settings, while conventional pathways may be limited by geography, cost, waiting time, stigma, and the difficulty of monitoring symptoms between appointments. Artificial intelligence is therefore attractive not only because it can classify complex observations, but because it can distribute selected functions across time and place. The clinical value of that distribution, however, is not determined by predictive accuracy alone. A technically capable classifier can remain distant from a treatment decision, and a highly accessible intervention can remain weakly supported if its evaluation lacks comparison, sustained observation, or a clinically interpretable endpoint.

The distinction between computational performance and clinical usefulness has become increasingly important as machine learning has spread across mental-health research. Broad surveys have documented uses in diagnosis, prognosis, treatment selection, symptom monitoring, and service administration.[2,3,24] The underlying learning paradigm replaces manually specified decision rules with patterns estimated from observations.[23] Deep neural networks can learn nonlinear representations from images, language, and longitudinal records,[6,8] while mobile sensing can transform everyday device traces into estimates of mood, sleep, mobility, and social behaviour.[7,38,40] Related work has examined emotion-sensitive robotics and technology-supported dementia care.[4,9] These capabilities enlarge the observational field available to clinicians, yet they also place multiple inferential steps between a recorded signal and a therapeutic consequence. Device use may correlate with distress without identifying its cause; a language model may quantify therapist behaviour without showing that altering that behaviour improves symptoms; and a neuroimaging classifier may separate groups without demonstrating transportability to routine clinics.

Diagnostic applications illustrate the tension. Neuroimaging studies of attention-deficit disorders, schizophrenia, cognitive impairment, and treatment response have reported promising discrimination or prediction.[10,13,16,28] Clinical text has likewise supported depression phenotyping and outcome research.[17,36] These studies address information that is difficult to interpret manually and may reveal patterns distributed over thousands of features. Their clinical path is nevertheless long: a model must survive acquisition differences, population shift, missingness, and changes in prevalence before it can inform an individual decision. Reviews of psychiatric neuroimaging have consequently emphasized methodological heterogeneity and the need for external validation rather than accepting a single accuracy estimate as sufficient evidence.[13,14] In suicide prevention, the same caution is amplified because false reassurance and excessive alerting can both cause harm.[26,27,37]

Intervention systems occupy a different position. Conversational agents can deliver elements of cognitive behavioural therapy, psychoeducation, or supportive dialogue at any hour; online programs can structure insomnia treatment; virtual reality can support exposure or mindfulness practice; and telephone or internet delivery can reduce travel requirements. Randomized evaluations of conversational treatment have reported symptom reductions among young adults, and related work has examined psychological agents for depression and anxiety.[15,29] Online cognitive behavioural therapy has been compared with clinic delivery in adolescents,[20] while automated cognitive behavioural therapy for insomnia has been studied as a route to depression prevention.[19] Virtual environments can create controllable therapeutic experiences that would be difficult to arrange in everyday settings.[18,33,35] These technologies are close to the patient, but proximity itself creates obligations: crisis responses, escalation pathways, privacy protections, comprehensible consent, and clear boundaries between automated interaction and licensed care are essential.[11,32,34]

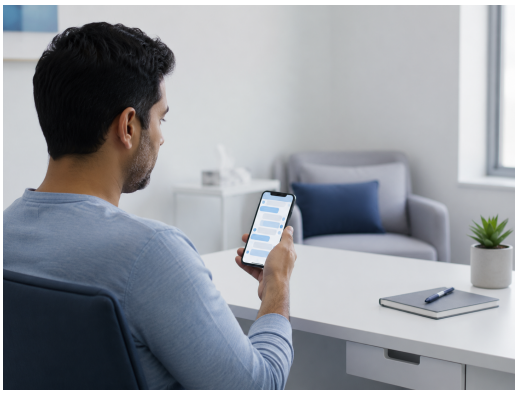
Another group of systems operates between diagnosis and intervention. Passive sensing may identify behavioural changes before a scheduled visit, and psychotherapy-language analysis may characterize treatment content at a scale unavailable to human raters. Personal sensing has been described as a layered translation from raw signals to behaviourally meaningful states.[7] Large clinical-language analyses have connected session content with outcomes across thousands of patients.[21] These systems may improve supervision, measurement, or timing without themselves delivering therapy. Treating them as equivalent to a patient-facing intervention obscures this difference in clinical action. Conversely, excluding them from discussions of care would ignore their potential influence on triage, clinician attention, and quality assurance.

The evaluation problem is therefore multidimensional. Population size affects precision and the opportunity to examine heterogeneity, but a large retrospective cohort does not provide the same causal leverage as a randomized comparison. Study design affects attribution, but a small randomized trial may still yield an unstable effect estimate. Patient contact affects implementation relevance, but direct access can magnify both benefit and harm. Endpoint proximity affects interpretation: symptom change, relapse, or functional improvement bears a different relation to care than a process label, a latent representation, or a classification output. Each property is meaningful, and none can substitute completely for the others. A defensible assessment should show how conclusions change when their relative importance is

altered.

Existing discussions often organize the field by algorithm, disorder, or delivery channel, and psychotherapy scholarship has separately asked how computational tools could change treatment research and practice.[12] Algorithm-centred organization clarifies technical capability but can conceal the distance between prediction and treatment. Disorder-centred organization clarifies clinical population but mixes direct interventions with observational tools. Delivery-centred organization clarifies accessibility but may overlook study design. The present study instead asks: among early artificial-intelligence applications in psychological services, which combination of population scale, design strength, care proximity, and endpoint proximity produces a stable evidentiary ordering, and does the apparent advantage of patient-facing delivery remain when the care criterion is removed from the statistical comparison? This question differs from asking whether artificial intelligence is generally promising. It concerns the conditions under which a clinical claim remains persuasive when reasonable evaluative priorities vary.

To answer it, nine investigations were converted into a transparent study-level matrix. The analysis did not assign technical accuracy values where none were available, combine unlike symptom scales, or impute clinical effects. It evaluated observable design properties: exact population size, comparative leverage, distance from direct care, and distance from symptom or treatment outcomes. Unreported population totals were treated probabilistically, criterion weights were perturbed 100,000 times, and each criterion was removed in turn. An exact permutation analysis compared patient-facing interventions with diagnosis and measurement applications using a score that deliberately excluded care proximity. The emphasis is consequently on robustness and auditability rather than a pooled treatment effect. This approach permits a precise answer from heterogeneous evidence without pretending that classification accuracy, acceptability, symptom reduction, and therapy-process measurement share a common clinical unit.



(a) Conversational CBT

(b) Neuroimaging review

(c) Guided VR practice

Figure 1. Clinical contact across delivery modes.

The care environments in Fig. 1 make the analytic distinction concrete. Conversational and virtual-reality systems interact with a patient during treatment, whereas the neuroimaging application supports a professional judgement before treatment. The images are illustrative depictions rather than photographs of participants in the analysed investigations.

2. MATERIALS AND METHODS

2.1. Study identification and eligibility

The retrieval phase used PubMed and Web of Science to identify English-language articles dated from January 2010 through December 2020. Search terms combined “psychological interventions,” “diagnosis on mental health disorders,” “artificial intelligence,” and “deep learning.” Eligibility required clinical, genetic, or medical information from patient populations. Nine records met these conditions and formed the analytic cohort.[1] The retained set covered an automated conversational agent, personal sensing, functional magnetic resonance imag-

ing classification, depression phenotyping from clinical notes, virtual-reality dialectical behaviour therapy, automated online cognitive behavioural therapy for insomnia, online versus clinic cognitive behavioural therapy, neural analysis of psychotherapy transcripts, and a meta-analysis of telephone psychotherapy.

The unit of analysis was an investigation, not an individual participant. Exact participant totals were recorded when the study table supplied a number. Session count was kept separate from patient count for the psychotherapy-language study because repeated sessions do not constitute independent patients. The sensing article described a layered model rather than a participant total, and the telephone-psychotherapy article described eligible randomized trials without a single combined total. Those two entries were marked as unreported rather than assigned an invented count.

2.2. Study-level coding

Four criteria were coded before numerical analysis. Population scale represented the amount of participant-level observation and was calculated only from explicit counts. Design strength represented the degree of comparative leverage visible from the article description: randomized controlled trials and a meta-analysis restricted to randomized trials received 1.00; a predefined randomized protocol received 0.75; cohort, classification, feasibility, and retrospective language analyses received 0.50; and a methodological sensing model received 0.25. These values distinguish causal contrast from descriptive learning without claiming that every design within a category has identical quality.

Care proximity represented the system's distance from direct treatment. A value of 1.00 denoted delivery of, or direct comparison between, psychological treatments; 0.50 denoted measurement embedded near care, such as personal sensing or psychotherapy-language analysis; and 0 denoted diagnosis-only classification. Endpoint proximity represented the clinical meaning of the reported target. Symptom change, treatment response, prevention, or a direct psychotherapy comparison received 1.00; diagnosis or patient-level prediction received 0.75; and behaviour-state or therapy-process measurement received 0.50. The coding deliberately avoided algorithm accuracy because comparable values were not available across the nine records.

The complete analytic matrix appears in Table 1. The categories are intentionally coarse enough to be auditable from the article descriptions, yet sufficiently discriminating to separate a patient-facing trial from a technical classification exercise. Ambiguous information was not upgraded on the basis of presumed study quality. In particular, the absence of a participant total remained explicit throughout the uncertainty analysis.

The cohort is heterogeneous by design, but its heterogeneity is clinically interpretable. Five records concern direct treatment delivery, two concern diagnosis, and two concern measurement near treatment. Seven records report exact population counts. The coding therefore preserves the distinctions that a pooled effect estimate would erase, while providing a common structure for assessing how much each distinction affects the ordering.

2.3. Clinical evidence proximity score

For an investigation i, population scale S_i was compressed logarithmically because the observed counts spanned more than three orders of magnitude:

$$S_i = \frac{\log_{10}(N_i)}{\log_{10}(17{,}572)}. \tag{1}$$

The denominator is the largest reported participant count in the cohort. Logarithmic compression prevents the 17,572-patient record from overwhelming all smaller trials while retaining the ordering of population sizes. It also recognizes that increasing a study from 15 to 150 participants changes evidentiary precision more substantially than an equal absolute increase at the upper end of the distribution.

The clinical evidence proximity score C_i combined population scale, design strength D_i, care proximity P_i, and endpoint proximity E_i:

$$C_i = 100\,(0.30S_i + 0.30D_i + 0.20P_i + 0.20E_i)\,. \tag{2}$$

Population scale and design received slightly greater nominal weight because they govern precision and attribution. Care and endpoint proximity jointly accounted for the remaining 40%, ensuring that a large observational analysis could not be treated as equivalent to a direct trial solely because of size. The score is an ordinal decision aid rather than a clinical effect estimate. A difference of ten points indicates a different combination of study properties; it does not indicate ten percentage points of symptom improvement.

2.4. Weight perturbation and missing-count uncertainty

Reliance on one weight vector would make the ordering depend on a single value judgement. The four weights were therefore sampled from a Dirichlet distribution centred on $(0.30, 0.30, 0.20, 0.20)$ with concentration 80. This concentration allows meaningful movement while keeping extreme, clinically implausible allocations uncommon. For each of 100,000 draws, the score and rank of every investigation were recalculated.

The two unreported population totals were varied within the same experiment. Their scale components were independently sampled from a symmetric Beta$(2, 2)$ distribution on $[0, 1]$. This distribution places greater mass near the centre than the extremes but permits any position from the smallest to the largest normalized population. The procedure produces a 95% uncertainty interval for

Table 1. Analytic cohort and criterion coding.

Investigation	Clinical role	Delivery	N	Design	Care	Endpoint
Fitzpatrick et al.	Intervention	Conversational CBT	70	1.00	1.00	1.00
Mohr et al.	Measurement	Personal sensing	–	0.25	0.50	0.50
Kuang and He	Diagnosis	fMRI classification	449	0.50	0	0.75
Geraci et al.	Diagnosis	Clinical-note neural network	366	0.50	0	0.75
Navarro-Haro et al.	Intervention	VR–DBT mindfulness	44	0.50	1.00	1.00
Gosling et al.	Intervention	Automated CBT-I	1,600	0.75	1.00	1.00
Spence et al.	Intervention	Online versus clinic CBT	15	1.00	1.00	1.00
Ewbank et al.	Measurement	Psychotherapy transcripts	17,572	0.50	0.50	0.50
Castro et al.	Intervention	Telephone psychotherapy	–	1.00	1.00	1.00

CBT, cognitive behavioural therapy; CBT-I, cognitive behavioural therapy for insomnia; DBT, dialectical behaviour therapy; fMRI, functional magnetic resonance imaging; VR, virtual reality. The Ewbank record also included 90,934 therapy-session transcripts. A dash denotes an unreported combined participant total.

each score, the probability of finishing first, and the probability of finishing among the first three positions. A fixed seed of 20260814 makes all values reproducible.

2.5. Criterion omission and exact group comparison

Each criterion was removed in turn, and the remaining three were averaged with equal weights. This analysis tests whether a record's position depends on one favourable property. For example, removing population scale reveals whether a large cohort is masking limited clinical proximity, whereas removing design strength reveals whether a small randomized comparison remains prominent on the other attributes.

Patient-facing interventions were also compared with diagnosis and measurement applications. To avoid circularity, care proximity was excluded from the comparison score; population scale, design strength, and endpoint proximity were averaged equally. The observed mean difference was assessed by enumerating all $\binom{9}{5} = 126$ allocations of five records to the patient-facing group. The two-sided exact probability was the proportion of allocations with an absolute mean difference at least as large as the observed value. Given the small cohort, the result is interpreted as an internal contrast rather than a population-wide causal claim.

3. RESULTS

3.1. Composition and scale of the analytic cohort

The nine investigations comprised five patient-facing interventions (55.6%), two diagnostic applications (22.2%), and two measurement applications located near care (22.2%). Exact participant counts were available for seven records (77.8%). Those records represented 20,116 participants, with a median of 366 and a range from 15 to 17,572. The psychotherapy-language investigation contributed 17,572 patients and 90,934 sessions; the latter figure was not added to the participant total. Four records had an explicit randomized comparison, randomized protocol, or randomized-trial synthesis, whereas the remaining five were classification, feasibility, retrospective, or methodological studies.

Population size and clinical role were not aligned. The largest cohort concerned the language of psychotherapy sessions rather than direct symptom delivery. The second-largest cohort concerned automated insomnia treatment and depression prevention. At the other extreme, the 15-participant adolescent trial directly compared online and clinic cognitive behavioural therapy. Consequently, any ordering based only on sample size would place the treatment-process study first and the randomized adolescent comparison last, even though their inferential purposes differ substantially.

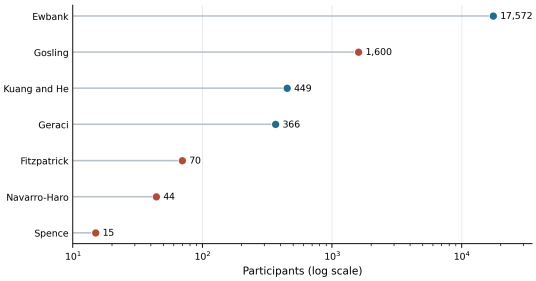


Figure 2. Reported participant counts.

The logarithmic display in Fig. 2 shows the concentration of five investigations below 500 participants and the separation of the two largest cohorts. Colour encodes clinical role rather than quality: the red points are direct interventions, while blue points denote diagnosis or measurement. The display also clarifies why logarithmic compression was required in Eq. (1); a linear axis would render most records visually indistinguishable near zero.

3.2. Nominal scores and uncertainty intervals

Nominal values and perturbation results are reported in Table 2. Under nominal weights and a midpoint value for unreported population scale, telephone psychotherapy had a score of 85.0, automated CBT-I 85.1, conversational

CBT 83.0, and online-versus-clinic CBT 78.3. The diagnostic records scored 48.7 and 48.1. Personal sensing had the lowest midpoint value, 42.5, because it provided neither a participant total nor a direct treatment endpoint in the analytic description.

The leading pair had similar median scores but different uncertainty. Automated CBT-I remained within 82.6–87.8 because every criterion was observed. Telephone psychotherapy ranged from 70.0 to 97.3 because its combined participant count was not reported. Thus, the latter attained first place slightly more often but was less precisely located. Conversational CBT rarely finished first, yet its 99.95% probability of a first-three position made it the most stable member of the leading group. Its exact population count and randomized design prevented the broad uncertainty seen for telephone psychotherapy.

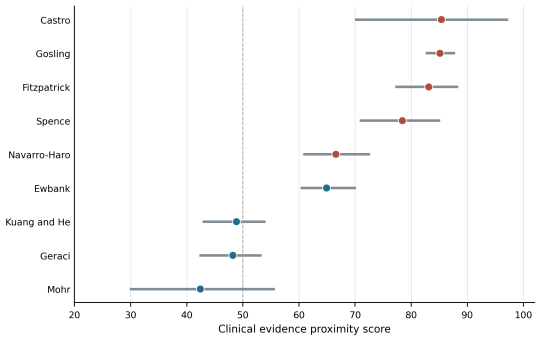


Figure 3. Score medians and 95% intervals.

The interval geometry in Fig. 3 separates two kinds of result. Records with complete population information have comparatively narrow intervals driven mainly by weight variation, whereas the two records without a combined participant total have visibly wider intervals. The telephone-psychotherapy median remains high across that uncertainty, but its exact position cannot be stated as precisely as the position of automated CBT-I.

The middle positions also reveal why population size cannot stand alone. The 17,572-patient psychotherapy-language study had a median of 64.9, below the 66.6 value for the 44-participant virtual-reality intervention. Its population component was maximal, but the record concerned treatment-process measurement and lacked a randomized therapeutic contrast. Conversely, the virtual-reality study was clinically proximal but limited by a small feasibility design. These two profiles reached similar total values through different strengths and should not be interpreted as interchangeable evidence.

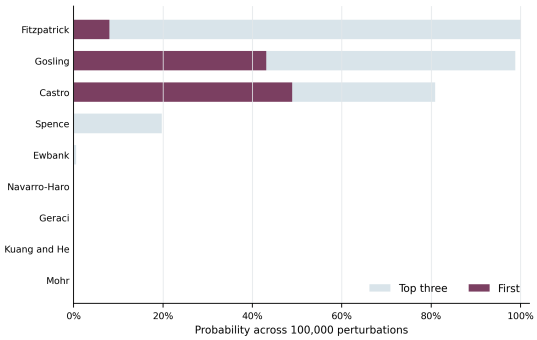


Figure 4. Rank attainment probabilities.

The nested bars in Fig. 4 distinguish occasional leadership from stable membership in the leading set. Telephone psychotherapy had the largest first-place probability, yet conversational CBT had the highest probability of remaining among the first three. Spence et al. entered that group in 19.7% of draws, showing that its strong randomized comparison became competitive when the population-size weight decreased.

3.3. Rank stability under criterion omission

Removing population scale produced a three-way tie at the top among conversational CBT, online-versus-clinic CBT, and telephone psychotherapy, each of which combined the strongest design and direct clinical endpoints. Automated CBT-I followed because its protocol-level design value was 0.75 rather than 1.00. This ordering demonstrates that the small sample of the adolescent CBT trial was the principal reason it did not enter the leading pair under nominal weighting.

Removing design strength placed automated CBT-I first, followed by telephone psychotherapy and conversational CBT. Removing care proximity still left the patient-facing records near the top because their endpoint values and comparative designs remained strong. Removing endpoint proximity produced a similar pattern, although the diagnosis-only records fell because neither care nor symptom delivery compensated for moderate design values. The leading set was therefore not created by one criterion alone, but exact positions within that set depended on whether sample scale or design received emphasis.

The psychotherapy-language record remained near the middle under every omission. Population scale secured a strong contribution, but removing that criterion reduced it to the same care and endpoint values as other measurement applications. The sensing record remained last or near last because its participant count was unreported and its description concerned a translation model rather than a comparative clinical evaluation. These findings distinguish stable moderation from instability: the language study was consistently intermediate, whereas the telephone-psychotherapy synthesis moved widely because of missing population scale.

Table 2. Score robustness across 100,000 perturbations.

Investigation	Nominal	Median	95% interval	First	First three
Fitzpatrick et al.	83.0	83.1	77.2–88.4	8.0%	99.95%
Mohr et al.	42.5	42.4	29.8–55.7	0	0
Kuang and He	48.7	48.9	42.8–54.0	0	0
Geraci et al.	48.1	48.2	42.3–53.4	0	0
Navarro-Haro et al.	66.6	66.6	60.8–72.7	0	0
Gosling et al.	85.1	85.1	82.6–87.8	43.1%	98.8%
Spence et al.	78.3	78.4	70.8–85.2	0	19.7%
Ewbank et al.	65.0	64.9	60.3–70.2	0	0.6%
Castro et al.	85.0	85.4	70.0–97.3	48.9%	80.9%

The interval reflects simultaneous weight perturbation and uncertainty in the two unreported participant totals. Percentages are rank probabilities.

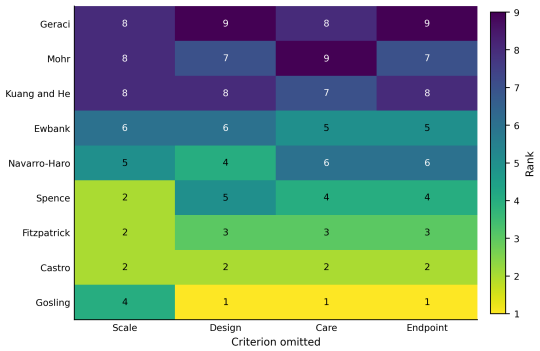


Figure 5. Ranks after single-criterion omission.

The heat map in Fig. 5 confirms that no leading record relied exclusively on care or endpoint proximity. When population scale was omitted, three strong comparative interventions shared the leading position; average ranks are printed for ties. Automated CBT-I led under the other three omissions, demonstrating the consistency produced by its observed population, direct delivery, clinical endpoint, and predefined randomized design.

3.4. Patient-facing versus analytic applications

On the care-excluded score, patient-facing interventions averaged 77.36 and the diagnosis/measurement records averaged 58.16, a difference of 19.20 points. Only three of the 126 possible allocations produced an absolute difference at least this large, giving a two-sided exact probability of $p = 0.0238$. Because care proximity was absent from the calculation, the difference arose from stronger endpoint connection and, for several intervention records, stronger comparative designs rather than from assigning the intervention group a care score of 1.00.

The comparison does not establish that an intervention algorithm is inherently superior to a diagnostic or measurement algorithm. The criteria address proximity of evidence to clinical action, not technical discrimination or biological discovery. A diagnosis-only study may be scientifically valuable while receiving a lower clinical proximity score. The exact test instead shows that, within this cohort, direct-care records also tended to possess study features that supported a more immediate clinical interpretation.

4. DISCUSSION

This study asked which properties govern a stable evidentiary ordering among early artificial-intelligence applications in psychological services. The answer is not population size, algorithm family, or delivery channel in isolation. Stable leadership arose when direct clinical contact was coupled with comparative leverage and an endpoint tied to treatment or symptoms. Automated CBT-I, telephone psychotherapy, and conversational CBT formed the leading group across weight perturbations, but for different reasons. Automated CBT-I combined a large observed population with direct delivery and a predefined randomized design. Telephone psychotherapy combined randomized-trial synthesis with direct symptom relevance but lacked a single auditable participant total. Conversational CBT combined an exact population count with a randomized design and direct symptom delivery, producing exceptional first-three stability even though its smaller sample reduced the probability of ranking first.

The result clarifies a recurrent problem in digital mental-health evaluation. Accessibility is often presented as clinical value, yet access does not itself demonstrate benefit. A system may be available continuously and still lack an adequate comparator, an interpretable endpoint, or a safe escalation procedure. Conversely, rigorous clinical design cannot guarantee implementation value if the intervention is unusable outside a controlled setting. Research on conversational agents has shown both the appeal of immediate support and the need to distinguish structured cognitive behavioural content from open-ended human therapy.[11,15,29] The present ordering reflects that distinction by rewarding direct delivery without allowing it to dominate design and population information.

The findings also explain why a very large clinical-language cohort occupied an intermediate position. Neu-

ral analysis of 90,934 psychotherapy sessions from 17,572 patients creates an unusually rich view of treatment process.[21] Such scale can reveal associations that a manually rated sample would miss, and automated coding may improve supervision or fidelity assessment. Yet the clinical claim is one inferential step removed from changing care. Association between language patterns and outcomes does not show that deliberately changing those patterns will improve an individual's symptoms. A prospective test of language-guided feedback would be required to close that step. The score therefore treats large-scale process analysis as valuable but not equivalent to a randomized treatment comparison.

Diagnostic records occupied lower positions for a related reason. Neuroimaging and clinical-text classifiers can improve phenotyping and may eventually guide treatment selection.[16,17,28] Their immediate endpoint, however, is classification or prediction. Before deployment, developers must establish calibration in the target clinic, robustness to acquisition differences, incremental value over clinician judgement, and consequences of false decisions. Reviews of neuroimaging and electroencephalography have noted substantial variation in preprocessing, validation, and population composition.[13,14] A clinically credible pathway therefore includes more than discrimination in a development cohort. The lower proximity scores do not reject diagnostic research; they identify the additional evidence needed to connect a technical output to improved care.

The exact group comparison strengthens this interpretation. Care proximity was removed before patient-facing records were compared with the remaining records, so the 19.2-point difference cannot be attributed mechanically to the group definition. Direct-care investigations tended also to use randomized contrasts and symptom-linked endpoints. This association may partly reflect disciplinary convention: psychotherapy research has long relied on controlled comparisons and symptom scales, while machine-learning research often emphasizes prediction on a held-out sample. Bringing those traditions together would improve both. Diagnostic developers can incorporate decision-curve analysis, prospective evaluation, and patient-centred outcomes; intervention researchers can strengthen external validation, subgroup assessment, and monitoring for distributional change.

Population-scale compression was consequential. Without logarithmic transformation, the 17,572-patient record would dominate any linear weighted score and obscure the evidentiary contribution of smaller randomized trials. The logarithmic form acknowledges diminishing returns in relative information while retaining a reward for larger populations. Criterion omission showed that removing population scale elevated the 15-participant adolescent CBT comparison into a tie for first. This is not evidence that 15 participants are sufficient. It demonstrates precisely why population size must remain visible alongside design: the trial asked a clinically direct question with a strong comparison, but its estimate would be imprecise and vulnerable to idiosyncratic sampling. The multidimensional result preserves both facts.

Unreported participant totals created a different form of uncertainty. Telephone psychotherapy was ranked first in 48.9% of perturbations and among the first three in 80.9%, but its interval was much wider than that of automated CBT-I. Reporting quality thus affects not only reproducibility but decision stability. A synthesis may be methodologically strong, yet an absent combined count prevents readers from judging the breadth of evidence at a glance. The probabilistic treatment used here avoids two common errors: assigning zero, which would imply no participants, and assigning the cohort median, which would conceal uncertainty. The same principle applies to digital mental-health reporting more broadly. Missing attrition, engagement, follow-up, subgroup, or adverse-event information should remain visible rather than being silently interpreted as favourable.

Clinical endpoint proximity deserves particular attention because digital systems can optimize what is easy to measure. Clicks, messages, linguistic markers, sensor traces, and completion rates are abundant. Symptom remission, functional recovery, therapeutic alliance, crisis avoidance, and sustained wellbeing are more difficult and costly to observe. A system trained on the former may improve its numerical objective without improving the latter. Work on personal sensing, mobile mood assessment, and behavioural distinctions in online communication demonstrates the richness of continuous traces,[7,25,38,40] but translation into clinical action requires explicit evidence that a timely response improves outcomes. The score's endpoint component makes this requirement visible without dismissing intermediate measurements.

Human responsibility remains central across all three clinical roles. Automated agents should disclose their non-human status, define the limits of advice, protect sensitive conversation, and route acute risk to qualified services. Studies of commercial conversational agents have shown uneven responses to questions involving mental health and interpersonal violence.[34] Ethical analyses have also emphasized privacy, dependence, deception, and the redistribution of professional responsibility.[11] For passive sensing, valid consent is complicated by continuous collection and by inferences that users may not anticipate. Consent requires particular care when a recommendation is generated from opaque behavioural patterns. Systems that simulate social understanding also raise questions about what users infer concerning machine comprehension and reciprocity.[39] These concerns are not optional implementation details; they affect whether a clinically proximal system can be used responsibly.

Virtual and remote interventions illustrate how modality can alter both therapeutic opportunity and evidentiary need. Virtual reality offers controlled exposure and

attentional practice, and it may increase vividness or engagement.[18,35] The 44-participant mindfulness study received strong care and endpoint values but only moderate design strength, placing it below the leading randomized records. A larger controlled trial with symptom follow-up would improve its position more effectively than a more elaborate rendering engine. Telephone psychotherapy, by contrast, uses technically simple delivery but carries strong clinical evidence from randomized comparisons.[22] This contrast is important: technological novelty and clinical readiness are not synonymous.

Treatment selection is another promising but demanding use. Neuroimaging and deep learning have been used to predict antidepressant response,[30,31] while broader psychiatric work has explored neural networks for stratification and prognosis.[36] A selection model should ultimately be tested through treatment assignment, not merely correlation with response under existing practice. That evaluation must compare outcomes when decisions are model-assisted with outcomes under an appropriate clinical alternative. It should also assess calibration across demographic and clinical subgroups, because an average gain can conceal harmful performance differences. The design criterion would rise only when that prospective comparison is present.

Several limitations bound the interpretation. First, the cohort contains nine records and is restricted to articles selected for positive findings. The distribution of scores therefore cannot estimate the field-wide prevalence of strong evidence and may overstate readiness. Second, coding was limited to information available in the study descriptions and article titles. It did not recover unreported attrition, allocation concealment, diagnostic prevalence, demographic composition, or adverse events. Third, the score weights express a clinical ordering of priorities rather than a universal law. The perturbation experiment shows whether conclusions survive reasonable movement around those priorities, but another decision context may justifiably place greater weight on safety, equity, cost, or external validation.

Fourth, the analysis does not combine clinical effects. The records address unlike questions, including classification, feasibility, process measurement, symptom change, and evidence synthesis. A standardized treatment effect would be meaningful only for a more homogeneous subset. Fifth, the endpoint scale is ordinal. A value of 1.00 indicates a direct clinical endpoint, not proof that the observed change was large, durable, or unbiased. Sixth, the search window ended in 2020. The analysis consequently describes an early clinical evidence structure and should not be read as an inventory of current products. These limitations narrow the claim appropriately: the study identifies what made evidence more clinically interpretable within the eligible cohort, not which technology should be purchased or prescribed today.

The results suggest a practical research sequence. Developers should define the clinical decision or treatment action before choosing an algorithm; select an endpoint that reflects patient benefit; state the comparator; preserve an auditable participant flow; and report uncertainty, attrition, and adverse events. Diagnostic models should be tested prospectively at the point of decision. Measurement systems should be evaluated through a trial of the action they trigger. Patient-facing interventions should combine controlled clinical evaluation with escalation pathways and post-deployment surveillance. These requirements align technical development with the consequences that matter to patients and clinicians.

5. CONCLUSION

The paper's research question concerned which study properties produce a robust ordering of clinical evidence proximity and whether patient-facing delivery retains an advantage after care proximity is removed from the comparison. Across 100,000 perturbations, direct interventions occupied the leading positions when they also had comparative designs and symptom-linked endpoints. The patient-facing group remained 19.2 points higher on a score that excluded care proximity, with an exact probability of 0.0238. Population size improved stability but did not determine leadership: the largest cohort remained intermediate because it measured psychotherapy process, while smaller randomized treatment studies ranked higher because their conclusions were closer to clinical action. The most defensible conclusion is therefore conditional. Artificial intelligence becomes evidentially persuasive in psychological services when direct care relevance, comparative leverage, and clinically meaningful outcomes occur together and are documented with auditable population information. No single technical capability substitutes for that conjunction.

BIBLIOGRAFIA / REFERENCES

1. S. Zhou, J. Zhao, and L. Zhang, "Application of artificial intelligence on psychological interventions and diagnosis: An overview," *Frontiers in Psychiatry*, vol. 13, art. 811665, 2022. doi:10.3389/fpsyt.2022.811665.
2. S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H. C. Kim, et al., "Artificial intelligence for mental health and mental illnesses: An overview," *Current Psychiatry Reports*, vol. 21, art. 116, 2019. doi:10.1007/s11920-019-1094-0.
3. C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: A scoping review," *Translational Psychiatry*, vol. 10, art. 116, 2020. doi:10.1038/s41398-020-0780-3.
4. R. Savery and G. Weinberg, "A survey of robotics and emotion: Classifications and models of emotional interaction," arXiv:2007.14838, 2020. doi:10.48550/arXiv.2007.14838.
5. D. D. Luxton, "Artificial intelligence in psychological practice: Current and future applications and implications," *Professional Psychology: Re-*

search and Practice, vol. 45, pp. 332–339, 2014. doi:10.1037/a0034559.
6. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. doi:10.1016/j.neunet.2014.09.003.
7. D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annual Review of Clinical Psychology*, vol. 13, pp. 23–47, 2017. doi:10.1146/annurev-clinpsy-032816-044949.
8. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from electronic health records," *Scientific Reports*, vol. 6, art. 26094, 2016. doi:10.1038/srep26094.
9. H. Rostill, R. Nilforooshan, A. Morgan, P. Barnaghi, and T. Chrysanthaki, "Technology integrated health management for dementia," *British Journal of Community Nursing*, vol. 23, pp. 502–508, 2018. doi:10.12968/bjcn.2018.23.10.502.
10. M. Ismail, G. Barnes, M. Nitzken, A. Switala, and A. El-Baz, "A new deep-learning approach for early detection of shape variations in autism using structural MRI," in *2017 IEEE International Conference on Image Processing*, pp. 1057–1061, 2017. doi:10.1109/ICIP.2017.8296443.
11. A. Fiske, P. Henningsen, and A. Buyx, "Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy," *Journal of Medical Internet Research*, vol. 21, e13216, 2019. doi:10.2196/13216.
12. R. L. Horn and J. R. Weisz, "Can artificial intelligence improve psychotherapy research and practice?" *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 47, pp. 852–855, 2020. doi:10.1007/s10488-020-01056-9.
13. M. Quaak, L. van de Mortel, R. M. Thomas, and G. van Wingen, "Deep learning applications for the classification of psychiatric disorders using neuroimaging: Systematic review and meta-analysis," *NeuroImage: Clinical*, vol. 30, art. 102584, 2021. doi:10.1016/j.nicl.2021.102584.
14. C. Mateo de Bardeci, T. Cheng, and C. Soab, "Deep learning applied to electroencephalogram data in mental disorders: A systematic review," *Biological Psychology*, vol. 162, art. 108117, 2021. doi:10.1016/j.biopsycho.2021.108117.
15. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent: A randomized controlled trial," *JMIR Mental Health*, vol. 4, e19, 2017. doi:10.2196/mental.7785.
16. D. Kuang and L. He, "Classification on ADHD with deep learning," in *Proceedings of the International Conference on Cloud Computing and Big Data*, pp. 27–32, 2014. doi:10.1109/CCBD.2014.42.
17. J. Geraci, P. Wilansky, V. de Luca, A. Roy, J. L. Kennedy, and J. Strauss, "Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression," *Evidence-Based Mental Health*, vol. 20, pp. 83–87, 2017. doi:10.1136/eb-2017-102688.
18. V. Navarro-Haro, Y. López-del-Hoyo, D. Campos, M. M. Linehan, H. G. Hoffman, and A. García-Palacios, et al., "Meditation experts try virtual reality mindfulness: A pilot evaluation," *PLOS ONE*, vol. 12, e0187777, 2017. doi:10.1371/journal.pone.0187777.
19. J. A. Gosling, N. Glozier, K. Griffiths, L. Ritterband, F. Thorndike, and A. Mackinnon, et al., "The GoodNight study—online CBT for insomnia for the indicated prevention of depression: Study protocol for a randomised controlled trial," *Trials*, vol. 15, art. 56, 2014. doi:10.1186/1745-6215-15-56.
20. S. H. Spence, C. L. Donovan, S. March, A. Gamble, R. E. Anderson, and A. Prosser, et al., "A randomized controlled trial of online versus clinic-based CBT for adolescent anxiety," *Journal of Consulting and Clinical Psychology*, vol. 79, pp. 629–642, 2011. doi:10.1037/a0024512.
21. M. P. Ewbank, C. Cummins, T. Tablan, S. Bateup, A. Catarino, and M. A. J. Martin, et al., "Quantifying the association between psychotherapy content and clinical outcomes using deep learning," *JAMA Psychiatry*, vol. 77, pp. 35–43, 2020. doi:10.1001/jamapsychiatry.2019.2664.
22. A. Castro, M. Gili, I. Ricci-Cabello, M. Roca, and D. McMillan, "Effectiveness and adherence of telephone-administered psychotherapy for depression: A systematic review and meta-analysis," *Journal of Affective Disorders*, vol. 260, pp. 514–526, 2019. doi:10.1016/j.jad.2019.09.023.
23. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, pp. 255–260, 2015. doi:10.1126/science.aaa8415.
24. A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, pp. 1426–1448, 2019. doi:10.1017/S0033291719000151.
25. H. Harikumar, T. Nguyen, S. Gupta, S. Rana, R. Kaimal, and S. Venkatesh, "Understanding behavioral differences between short- and long-term drinking abstainers from social media," in *Advanced Data Mining and Applications*, pp. 520–533, 2016.
26. R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, and F. Abnousi, "Artificial intelligence and suicide prevention: A systematic review of machine learning investigations," *International Journal of Environmental Research and Public Health*, vol. 17, art. 5929, 2020. doi:10.3390/ijerph17165929.
27. L. Zheng, O. Wang, S. Hao, C. Ye, M. Liu, and M. Xia, et al., "Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records," *Translational Psychiatry*, vol. 10, art. 72, 2020. doi:10.1038/s41398-020-0684-2.
28. S. V. Kalmady, R. Greiner, R. Agrawal, V. Shivakumar, J. C. Narayanaswamy, and M. R. G.

Brown, et al., "Improving schizophrenia prediction with multiple brain parcellation ensemble learning," *npj Schizophrenia*, vol. 5, art. 2, 2019. doi:10.1038/s41537-018-0070-8.
29. R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, and M. Rauws, "Using psychological artificial intelligence to relieve symptoms of depression and anxiety: Randomized controlled trial," *JMIR Mental Health*, vol. 5, e64, 2018. doi:10.2196/mental.9782.
30. K. P. Nguyen, C. C. Fatt, A. Treacher, C. Mellema, and A. Montillo, "Predicting response to the antidepressant bupropion using pretreatment fMRI," in *Predictive Intelligence in Medicine*, 2019.
31. L. Squarcina, F. M. Villa, M. Nobile, E. Grisan, and P. Brambilla, "Deep learning for the prediction of treatment response in depression," *Journal of Affective Disorders*, vol. 281, pp. 618–622, 2020. doi:10.1016/j.jad.2020.11.104.
32. S. D'Alfonso, O. Santesteban-Echarri, S. Rice, G. Wadley, R. Lederman, and C. Miles, et al., "Intelligent automated online social therapy for youth mental health," *Frontiers in Psychology*, vol. 8, art. 796, 2017. doi:10.3389/fpsyg.2017.00796.
33. M. L. Tielman, M. A. Neerincx, R. Bidarra, B. Kybartas, and W.-P. Brinkman, "A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories," *Journal of Medical Systems*, vol. 41, art. 125, 2017. doi:10.1007/s10916-017-0771-y.
34. A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian, and E. Linos, "Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health," *JAMA Internal Medicine*, vol. 176, p. 719, 2016. doi:10.1001/jamainternmed.2016.0400.
35. C. Botella, J. Fernández-Álvarez, V. Guillén, A. García-Palacios, and R. Baños, "Recent progress in virtual reality exposure therapy for phobias: A systematic review," *Current Psychiatry Reports*, vol. 19, art. 42, 2017. doi:10.1007/s11920-017-0788-4.
36. D. Durstewitz, G. Koppe, and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Molecular Psychiatry*, vol. 24, pp. 1583–1598, 2019. doi:10.1038/s41380-019-0365-9.
37. T. M. Fonseka, V. Bhat, and S. H. Kennedy, "The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors," *Australian & New Zealand Journal of Psychiatry*, vol. 53, pp. 954–964, 2019. doi:10.1177/0004867419864428.
38. F. Gravenhorst, A. Muaremi, J. Bardram, A. Gruenerbl, O. Mayora, and G. Wurzer, et al., "Mobile phones as medical devices in mental disorder treatment: An overview," *Personal and Ubiquitous Computing*, vol. 19, pp. 335–353, 2015. doi:10.1007/s00779-014-0829-5.
39. F. Cuzzolin, A. Morelli, B. Cîrstea, and B. J. Sahakian, "Knowing me, knowing you: Theory of mind in AI," *Psychological Medicine*, vol. 50, pp. 1057–1061, 2020. doi:10.1017/S0033291720000835.
40. R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "MoodScope: Building a mood sensor from smartphone usage patterns," in *Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 389–402, 2013.