# From Measurement to Clinical Action: A Methodological Framework for ePROMs, Disease Modelling, and AI in Rural Mental Health Care


## Abstract

**Background:** Electronic patient-reported outcome measures (ePROMs) and artificial intelligence (AI) create opportunities to support longitudinal mental-health care. Their value depends, however, on the validity of what is measured, who can participate, how outputs are interpreted, and whether services can respond. These dependencies warrant particular attention in rural settings.

**Objective:** To propose an integrated methodological framework for guiding ePROM selection, design, disease modelling, interpretation, AI integration, and implementation in rural mental-health care.

**Approach:** A targeted narrative integration of measurement guidance, disease-course modelling methods, clinical prediction and AI standards, implementation frameworks, and selected rural mental-health studies informs a provisional conceptual framework. This is not a systematic review, consensus guideline, or empirical validation study.

**Proposed framework:** Ten connected domains link rural context, instrument selection, measurement design, data quality and longitudinal change, disease modelling, assessment of AI suitability, interpretation, clinical decision-making, implementation, and evaluation. Disease modelling addresses trajectories and illness-state transitions; its use does not require AI. Each domain requires a documented decision, supporting evidence, an accountable owner, and criteria for progression or revision. An optional multi-objective optimisation layer uses evolutionary algorithms to explore ePROM subsets, adaptive configurations, and monitoring choices under measurement, safety, and resource constraints. AI is optional: implementation should not proceed merely because a model is available. Equity, patient involvement, safety, and data governance apply across the pathway. Hypothetical service and compartmental examples illustrate non-AI decision-making and basic calculations without claiming empirical validation. A staged evaluation agenda separates the usability and decision quality of the framework from disease-model validity, predictive performance, and the effects of a resulting care intervention.

**Conclusion:** The proposed contribution is an auditable connection between measurement, analysis, and clinical action rather than a new instrument or algorithm. Participatory refinement and comparative evaluation are needed before claims of improved care, equity, or outcomes can be made.


**Keywords:** patient-reported outcome measures; rural mental health; disease modelling; artificial intelligence; multi-objective optimisation; measurement validity; health equity; implementation science.

# 1 Introduction

Digital mental-health care is often presented as a response to geographical barriers, but digital access and access to clinicians may be constrained in the same communities. A county-level US study found that limited broadband access co-occurred with lower availability of mental-health services; this ecological association does not establish the circumstances of every rural resident [1]. In an Australian rural care-navigation evaluation, service users described technical, privacy, and relational barriers to tele-mental health, while supportive navigation helped them consider available services [2]. These studies concern digital access and service delivery, not the effectiveness of AI-enabled ePROMs. They nevertheless identify contextual questions that such systems should address before deployment.

Patient-reported outcomes provide a way to bring patients' accounts into assessment and follow-up. Electronic collection can support repeated measurement, but digitising an instrument and making its scores clinically useful are different tasks. Guidance on electronic administration concerns whether changes of collection mode preserve measurement properties [3]; evidence on feedback-informed psychological treatment concerns what happens when information is incorporated into therapy. For example, a cluster-randomised trial found benefits of feedback technology for patients identified as at risk of poor treatment response [4]. That finding does not establish that any ePROM platform, prediction model, or rural implementation will produce comparable benefits.

Several established resources address different parts of this problem. COSMIN provides methods for evaluating evidence about measurement properties [5]. TRIPOD+AI addresses reporting of prediction-model studies, while PROBAST+AI addresses quality, risk of bias, and applicability [6, 7]. DECIDE-AI focuses on reporting early live clinical evaluations, and FUTURE-AI addresses trustworthy development and deployment more broadly [8, 9]. Implementation frameworks address organisational and contextual influences on adoption and sustainability [10, 11]. These resources are complementary, but none should be treated as interchangeable evidence that a particular local system is ready for clinical use.

Disease-course modelling adds a complementary methodological layer. Longitudinal models can represent repeated measurements, multi-state models can describe transitions between specified states, and joint models can connect repeated markers with event times [12–14]. These are modelling approaches, not evidence that ePROM-defined states correspond to clinical diagnoses or that an observed trajectory represents untreated illness. Their inclusion makes the representation of disease course explicit before considering AI.

Optimisation adds a further methodological question: which inputs and monitoring configurations offer acceptable trade-offs between useful information, patient burden, and service capacity? Genetic algorithms have been applied to personality-measure abbreviation, providing a methodological precedent rather than evidence of clinical validity for rural ePROMs [15]. Multi-objective evolutionary methods such as NSGA-II can search for competing solutions without reducing every objective to a single weighted score [16]. Their relevance depends on the decision problem; their use alone does not establish clinical benefit or methodological novelty.

The methodological problem considered here is how to connect these resources at the points where one decision constrains another. A well-performing model cannot resolve uncertainty about whether its input instrument measures the intended construct. An interpretable prediction does not establish that a feasible response exists. A high completion rate among digital users does not reveal whether

eligible non-users have been excluded. This article therefore treats the unit of interest as the whole measurement-to-action pathway, not the questionnaire or algorithm alone.

The aim is to propose a framework that makes these dependencies explicit for rural mental-health services. Its objectives are to connect instrument selection and design with longitudinal interpretation and disease modelling, define conditions for selecting or declining modelling, AI, and optimisation approaches, make equity and accountability operational, and specify how the framework itself could be evaluated. The proposed contribution is integration and decision traceability, not a claim to have discovered new measurement principles or to have produced a validated clinical guideline.

## 2 Scope and conceptual development

This is a conceptual and methodological article. The ten-domain framework spans rural context to outcome evaluation, distinguishes disease modelling from data preparation and AI suitability, and integrates selected literature narratively. Web-based searches were undertaken to locate original methodological publications, official guidance, and illustrative rural mental-health studies through publisher websites, PubMed records, and institutional repositories. Searches addressed rural mental health, patient-reported outcomes, disease modelling, optimisation, AI, measurement, and implementation. The following expanded Boolean blocks make these concepts explicit: quotation marks delimit individual terms or phrases, `OR` joins alternatives within a concept, and `AND` links different concepts. These blocks are illustrative formulations for a future search update, not a verbatim log of the searches already undertaken; syntax and phrase handling would need adaptation to the selected search platform.

**S1. Rural context:** ("rural" `OR` "remote" `OR` "underserved" `OR` "rural healthcare" `OR` "rural health services").

**S2. Mental health:** ("mental health" `OR` "mental healthcare" `OR` "mental illness" `OR` "depression" `OR` "anxiety" `OR` "psychological distress").

**S3. Patient-reported measurement:** ("ePROM" `OR` "ePROMs" `OR` "PROM" `OR` "PROMs" `OR` "patient-reported outcome" `OR` "patient-reported outcome measures" `OR` "electronic patient-reported outcomes").

**S4. Disease modelling:** ("disease modelling" `OR` "disease modeling" `OR` "disease trajectories" `OR` "symptom trajectories" `OR` "multi-state models" `OR` "multistate models" `OR` "joint models" `OR` "compartmental models" `OR` "model validation").

**S5. Evolutionary optimisation:** ("evolutionary algorithms" `OR` "genetic algorithms" `OR` "NSGA-II" `OR` "multi-objective optimisation" `OR` "multi-objective optimization" `OR` "item selection" `OR` "questionnaire optimisation" `OR` "adaptive measurement" `OR` "computerised adaptive testing" `OR` "computerized adaptive testing").

**S6. AI and prediction:** ("artificial intelligence" `OR` "AI" `OR` "machine learning" `OR` "prediction models" `OR` "longitudinal monitoring" `OR` "personalised care" `OR` "personalized care" `OR` "explainability").

**S7. Measurement and implementation:** ("measurement validity" `OR` "measurement comparability" `OR` "psychometric properties" `OR` "reliability" `OR` "missing data" `OR` "patient burden" `OR` "implementation" `OR` "health equity" `OR` "clinical safety").

**S8. Named methodological resources:** ("COSMIN" `OR` "TRIPOD+AI" `OR` "PROBAST+AI" `OR` "DECIDE-AI" `OR` "FUTURE-AI" `OR` "STANDING Together" `OR` "NASSS" `OR` "CFIR").

Illustrative combinations are `S1 AND S2 AND S3` for rural mental-health ePROMs; `S2 AND S3 AND (S4 OR S5 OR S6 OR S7)` for methodological applications; and `S8` for searches by resource name. Each block identifier denotes its full parenthesised expression, not a literal search term. Requiring every block simultaneously would unnecessarily restrict the methodological search. For example, one expanded expression is:

> ("rural" `OR` "remote" `OR` "underserved") `AND` ("mental health" `OR` "depression" `OR` "anxiety") `AND` ("ePROMs" `OR` "patient-reported outcome measures") `AND` ("genetic algorithms" `OR` "NSGA-II" `OR` "multi-objective optimisation" `OR` "multi-objective optimization").

Publisher or indexed bibliographic records were checked for the cited publications. Available full text and abstracts were used according to access.

Sources were selected purposively for their contribution to an identifiable decision in the pathway, rather than to estimate intervention effects or exhaustively catalogue the literature. The approach did not involve a registered review protocol, comprehensive database searching, duplicate screening, formal study-level quality assessment, or evidence grading. Consequently, the cited literature is illustrative rather than a complete evidence base, and no claim of systematic coverage or conceptual saturation is made. Appendix A makes the main source-to-domain links explicit.

The synthesis follows three analytical steps: separating measurement, disease-course representation, prediction, decision-support, and implementation questions; mapping relevant guidance to each domain; and specifying the decision records and dependencies needed to connect the domains. The decision gates, proposed documentation requirements, and evaluation hypotheses are conceptual proposals in this article, not findings attributed to the cited guidance. No patient data, interviews, Delphi exercise, psychometric testing, disease-model fitting, AI-model evaluation, or evolutionary-optimisation benchmarking were undertaken. The hypothetical application illustrates the proposed rules, not empirical feasibility. The compartmental calculations use entirely hypothetical inputs to demonstrate arithmetic; they are not estimates from rural patients.

For this article, an *ePROM* is a patient-reported outcome instrument administered electronically; the *platform* is the software and service arrangement through which responses are collected and used. Instrument design, electronic interface design, and care-pathway design are treated separately. *Disease modelling* means specifying and, where justified, estimating symptom trajectories, illness-state transitions, or time-to-event processes. It concerns disease course under stated observation and care conditions, not an assumed biological mechanism or an inevitable progression from mild to severe illness. *AI integration* principally concerns predictive or pattern-recognition models applied to longitudinal ePROM data, with conventional statistical models included as comparators. Autonomous diagnosis, automated treatment allocation, and generative conversational agents are outside the proposed initial scope. *Evolutionary optimisation* refers to a population-based search over candidate instruments, item subsets, or monitoring configurations; it is not synonymous with training a clinical predictor. *Personalisation* means adapting monitoring and shared care to an individual's circumstances and preferences, not assuming that a prognostic association identifies the treatment that will work best for that person.

The intended initial setting is adult outpatient or primary-care mental-health monitoring. Applying the framework to children, acute psychiatric care, or a different clinical purpose would require

separate specification and evaluation. Rurality should be defined using an appropriate jurisdiction-specific classification, supplemented by descriptions of travel, service availability, connectivity, language, and local resources. Rural populations should not be assumed to be uniformly underserved or culturally homogeneous.

## 3 Framework architecture and programme theory

Figure 1 presents ten connected domains. The ordering is a scaffold for deliberation, not a compulsory linear development process: selection and design inform each other, and evaluation may require revisiting any earlier decision. Domain 5 addresses disease-course representation and the suitability of a formal disease model; Domain 6 assesses whether AI is justified. A disease model can be statistical rather than AI-based, and AI may estimate components of a disease model rather than constitute a separate downstream model. Choosing neither formal modelling nor AI leaves a complete descriptive ePROM-based care pathway. The framework can guide appraisal of existing models and does not require developing a new model. Evolutionary optimisation is an optional method supporting selection and design in Domains 2–3 and configuration decisions in Domains 6–9, not an eleventh compulsory domain. The methodological contribution remains the specification, comparison, and evaluation of defensible choices rather than a commitment to one algorithm. All domains are subject to four cross-cutting commitments: meaningful patient and community involvement, equity, clinical safety, and accountable data governance. This architecture is consistent with treating technology-enabled care as a complex intervention whose operation depends on context and human activity [17].

The proposed programme theory is that explicit documentation of purpose, measurement evidence, data limitations, and response responsibilities will reduce unexamined transitions from scores to action. If patients can participate, measures capture relevant experiences, and staff have time and authority to respond, better-supported interpretation may improve shared decisions. Whether those decisions improve symptoms, functioning, quality of life, or equity remains a separate empirical question. The pathway may fail when increased measurement produces burden without action, when missing observations hide exclusion, or when model outputs displace rather than support clinical dialogue. These are hypotheses to investigate, not observed results of the framework.

Each domain produces an entry in a *decision record*: the intended use; alternatives considered; supporting and conflicting evidence; uncertainty and applicability limits; the decision and its justification; the responsible person or group; and the date or event that triggers review. Evidence should be identified as direct for the intended population, indirect from another context, or not yet available. This categorisation describes applicability rather than substituting for a formal certainty assessment. A service-level decision group should include clinical, measurement, implementation, and patient perspectives, with technical and community governance expertise where relevant.

Progression has three possible states: *proceed within a specified scope*, *revise or gather evidence*, or *do not implement the proposed component*. Meeting a reporting checklist is not, by itself, a progression criterion. Decisions require judgement proportionate to the intended use and potential harm. Unresolved critical safety or governance concerns cannot be averaged away by favourable ratings elsewhere.

A compact, proposed encoding of these decisions is

$$\mathcal{D}_k = \begin{cases} \text{do not implement,} & V_k = 1, \\ \text{revise or gather evidence,} & V_k = 0,\ A_k = 0, \\ \text{proceed within scope,} & V_k = 0,\ A_k = 1. \end{cases} \tag{1}$$

Here, $k$ denotes the domain or component under review, $V_k = 1$ indicates an unresolved critical safety or governance veto, and $A_k = 1$ indicates that the evidence and operational requirements for the specified use are judged adequate. Both indicators are documented judgements, not estimated probabilities. A veto prevents deployment of that component unless resolved; it need not stop an otherwise appropriate non-AI pathway. Equation (1) is a logical aid, not a validated score or a substitute for deliberation. It does not permit strong predictive performance to compensate for a critical safety failure.

# 4 Operational domains

## 4.1 Rural context and clinical purpose

The first decision is what problem requires better information and what a service could do differently if that information were available. A purpose statement should specify the population, care setting, intended users, decision, monitoring interval, and feasible response. Monitoring symptom trajectories before a scheduled review is different from predicting deterioration between appointments. Neither should be presented as continuous clinical surveillance unless the service actually supplies that function.

A proposed local assessment should document both constraints and assets: connectivity, private access to devices, digital and health literacy, travel demands, available clinicians, community organisations, trusted relationships, and options for assisted participation. Rural telehealth research indicates the importance of practical assistance and trust, but local arrangements must be established rather than inferred from another setting [2]. Patient involvement should include people who decline digital participation, not only confident platform users. Compensation, accessible meeting formats, and an explicit account of how contributions change decisions should be planned. The output is a context-and-purpose specification, not a technology procurement list.

## 4.2 ePROM selection

Selection begins with the construct and intended interpretation: symptom severity, functioning, quality of life, recovery, or another patient-prioritised outcome. Candidate instruments should be compared using evidence on content validity, relevant measurement properties, and feasibility. COSMIN emphasises whether items are relevant, sufficiently comprehensive, and understandable for the target construct, population, and context; reliability alone is not a sufficient selection argument [5].

The proposed selection record should explain what each instrument captures, what it omits, its population and language evidence, respondent burden, licensing conditions, and suitability for repeated use. A convenient short scale should not be preferred solely because it produces a complete

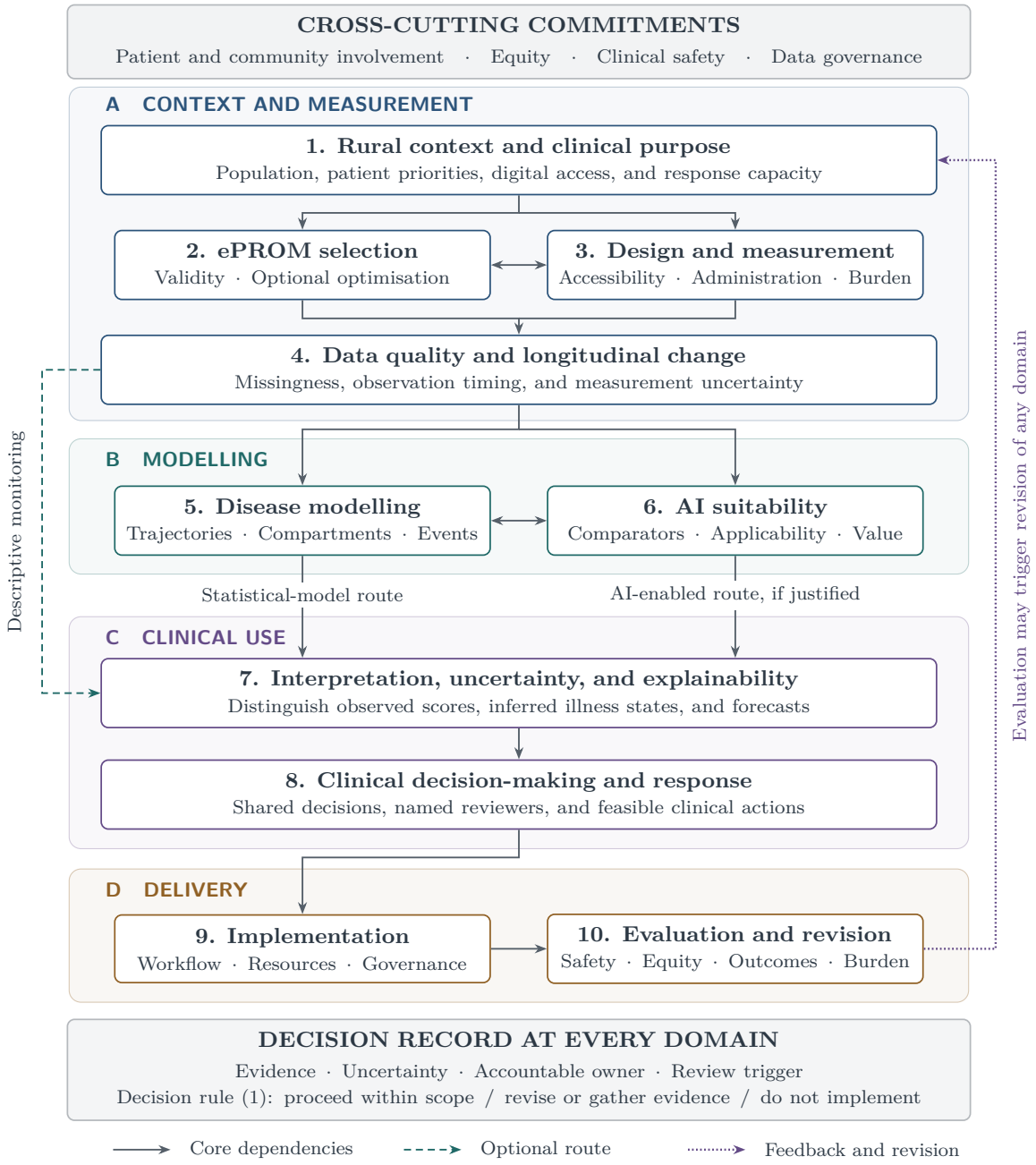



Figure 1: Layered measurement-to-action framework. Ten domains are grouped into context and measurement, modelling, clinical use, and delivery. Bidirectional links indicate iterative selection/design and coupled disease-modelling/AI choices. Evolutionary optimisation can support selection, design, and configuration without creating another mandatory domain. Statistical modelling need not use AI; descriptive monitoring remains an option after considering whether formal modelling or AI is warranted. These routes converge on interpretation and an actionable care response. The outer feedback loop may reopen any domain. Governance commitments and a decision record apply throughout. Arrows represent methodological dependencies and choices, not validated causal effects or mandatory model deployment.

dataset. Conversely, adding multiple scales requires a defensible clinical purpose. Where existing evidence is indirect, the team should specify the uncertainty and the work needed to address it, rather than describing an instrument as universally validated. A patient-prioritised goal can complement a standardised score without being treated as psychometrically interchangeable with that score. Instrument selection and interface design should be revisited together when accessibility requirements change.

Where a new instrument is justified because no suitable measure covers the intended construct, the proposed development sequence is concept elicitation with intended users, item generation, cognitive interviewing, and field testing of prespecified measurement properties. This is a distinct research programme, not an interface redesign or a prerequisite for every rural implementation. Its time and sample requirements should be compared with the option of using an existing measure within a narrower, defensible scope.

#### 4.2.1 Optional evolutionary optimisation of ePROM selection

The strongest initial application is choosing a clinically defensible measurement configuration while limiting unnecessary burden. An illustrative candidate pool might contain 50–100 items, but no such dataset or item bank is assumed to be available for this article. The primary comparison should be with an established instrument or approved short form. Selecting arbitrary items from validated questionnaires produces a new configuration whose validity cannot be inherited automatically. Where item-level modification is not permitted or cannot be supported, the search should operate over complete instruments or approved configurations instead. Calibrated item banks offer a principled alternative, with their own psychometric requirements [18].

Let $\mathbf{z} = (z_1, \ldots, z_p)$ encode inclusion of $p$ permitted candidate items, where $z_j \in \{0, 1\}$. For a broader monitoring-design problem, let $\Delta > 0$ be an administration interval in months and $\eta \in (0, 1)$ a threshold for an already specified routine-review prediction. The configuration is $\vartheta = (\mathbf{z}, \Delta, \eta)$; when only item selection is studied, $\Delta$ and $\eta$ remain fixed. A proposed multi-objective formulation is

$$\underset{\vartheta \in \Omega}{\text{minimise}}\ \mathbf{f}(\vartheta) = (\mathcal{L}_{\text{pred}}(\vartheta),\ B(\vartheta),\ R_{\text{miss}}(\vartheta),\ G(\vartheta))\,. \tag{2}$$

Here, $\mathcal{L}_{\text{pred}}$ is prediction loss estimated within development-data resampling for a prespecified target, $B$ is patient burden, $R_{\text{miss}}$ is estimated non-completion risk in the intended population, and $G$ is a prespecified disparity measure. For a purely measurement-focused application, prediction loss can be replaced by an appropriate measurement-error objective. The target must not change with the candidate subset merely to improve apparent performance. Reproducing a questionnaire total is not equivalent to predicting a patient-important clinical outcome.

Clinical relevance, construct coverage, and adequate measurement performance should not be sacrificed for a shorter questionnaire. An illustrative feasible set is

$$\Omega = \left\{ \vartheta : \begin{array}{l} \mathbf{z} \in \{0,1\}^p, \quad \Delta \in \mathcal{T}, \quad \eta \in \mathcal{A}, \\ \displaystyle\sum_{j=1}^{p} z_j \le K_{\max}, \quad \sum_{j \in \mathcal{J}_c} z_j \ge r_c \quad \text{for each required construct } c, \\ A_{\text{meas}}(\vartheta) = 1, \quad W(\vartheta) \le W_{\max}, \\ \mathcal{L}_g(\vartheta) \le U_g \quad \text{for each prespecified group } g \end{array} \right\}. \tag{3}$$

The sets $\mathcal{T}$ and $\mathcal{A}$ contain approved candidate intervals and routine-review thresholds. $K_{\max}$ limits item count, $\mathcal{J}_c$ identifies items covering construct $c$, and $r_c$ specifies required minimum coverage. These counts are screening constraints, not a complete content-validity assessment. $A_{\text{meas}} = 1$ denotes satisfaction of prespecified measurement and permissions requirements for research consideration, not authorisation for clinical deployment. These requirements should include suitable reliability or measurement precision for each intended score, rather than maximising a single internal-consistency coefficient across unrelated constructs. $W$ is anticipated clinician workload, bounded by locally agreed capacity $W_{\max}$. The group-specific loss limits $U_g$ are intended to prevent an aggregate improvement from concealing unacceptable performance for a subgroup; their uncertainty must be reported.

For example, if $c_j$ is an empirically estimated completion time in minutes, simple length and burden summaries are

$$K(\mathbf{z}) = \sum_{j=1}^{p} z_j, \qquad B(\vartheta) = \frac{1}{\Delta} \sum_{j=1}^{p} c_j z_j. \tag{4}$$

This approximation expresses response time per month and assumes additive item times; it omits setup time, distress, reminder fatigue, and assistance unless they are modelled separately. A possible disparity objective is $G(\vartheta) = \max_g |\mathcal{L}_g(\vartheta) - \mathcal{L}_{g_0}(\vartheta)|$ relative to group $g_0$. Reducing that gap by worsening care for everyone is not an acceptable definition of equity. Group loss limits, patient priorities, participation denominators, and the framework's wider equity assessment remain necessary. Unadministered items are planned omissions, not automatically missing responses.

#### 4.2.2 Evolutionary search and selection of trade-offs

A method such as NSGA-II can generate a population of candidate configurations, evaluate their objectives and constraints, and use selection, crossover, and mutation to explore alternatives. Its non-dominated sorting and diversity-preserving selection support a set of competing solutions rather than a single universally best subset [16]. In this application, candidates violating critical measurement, content, or safety requirements should not survive solely because they predict well.

For objectives that are all minimised, dominance and the ideal Pareto set can be written as

$$\begin{aligned} \vartheta' \prec \vartheta \iff f_r(\vartheta') \le f_r(\vartheta) \text{ for all } r, \quad f_s(\vartheta') < f_s(\vartheta) \text{ for some } s, \\ \mathcal{P} = \{\vartheta \in \Omega : \nexists \vartheta' \in \Omega \text{ with } \vartheta' \prec \vartheta\}. \end{aligned} \tag{5}$$

An evolutionary run generally provides an approximation based on noisy objective estimates, not proof of the global Pareto set. Patients and clinicians should choose among acceptable trade-offs using prospectively documented priorities; the algorithm must not silently determine how much patient burden is worth a change in prediction error. Figure 2 summarises this optional workflow.

Figure 3 illustrates binary item encoding and the genetic operators used in a possible NSGA-II implementation. The eight-item strings are schematic, not selected questionnaires. Single-point crossover and bit-flip mutation are illustrative choices for binary encoding, not requirements for every evolutionary algorithm. The lower panel depicts the parent–offspring survival mechanism of NSGA-II [16]; the clinical admissibility checks are requirements proposed by this framework.

The entire search, including item selection, tuning, imputation, and any data-driven trade-off selection, belongs inside the training process. Using the same cross-validation results to select a configuration and claim its final performance can introduce optimism; nested evaluation separates

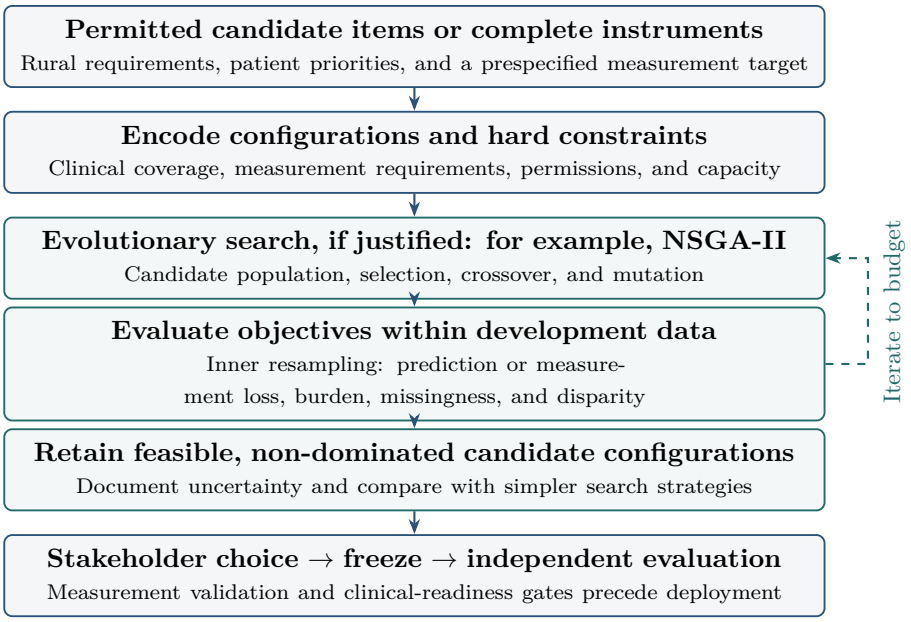



Figure 2: Optional evolutionary-optimisation workflow supporting ePROM selection and configuration. The output is a set of candidate trade-offs, not an automatically validated questionnaire or a guaranteed global optimum. Search and stakeholder selection use development information; independent evaluation must remain separate. This figure specifies a proposed method: no evolutionary search or clinical comparison has been performed for this article.

these tasks [19]. The proposed evaluation should preserve patient and temporal separation and use outer folds or an untouched external sample to assess the complete selection procedure. Where candidates change predictor inputs, the model class and tuning budget should be comparable and model fitting repeated within the appropriate training folds.

Evolutionary search should be compared with clinically selected short forms, greedy selection, and random search under comparable evaluation budgets; exact enumeration may be preferable when few configurations are feasible. Record seeds, population size, operators, stopping rules, objective definitions, constraint failures, computational cost, and stability across runs. If no added value is demonstrated, retain the simpler method. No evolutionary run, optimised item set, or performance advantage is reported here.

#### 4.2.3 Personalised ePROMs and monitoring configuration

Personalisation may involve choosing anxiety-relevant, depression-relevant, or treatment-response modules using the patient's preferences and prior responses. It must not mean that unrelated questionnaires are treated as interchangeable longitudinal measurements. Calibrated item banks and appropriate item-response models provide a basis for adaptive assessment, subject to requirements such as dimensionality, measurement precision, and examination of differential item functioning [18]. Evolutionary search can help design an admissible configuration or routing policy offline; it is not equivalent to the real-time item-selection rule of a computerised adaptive test.

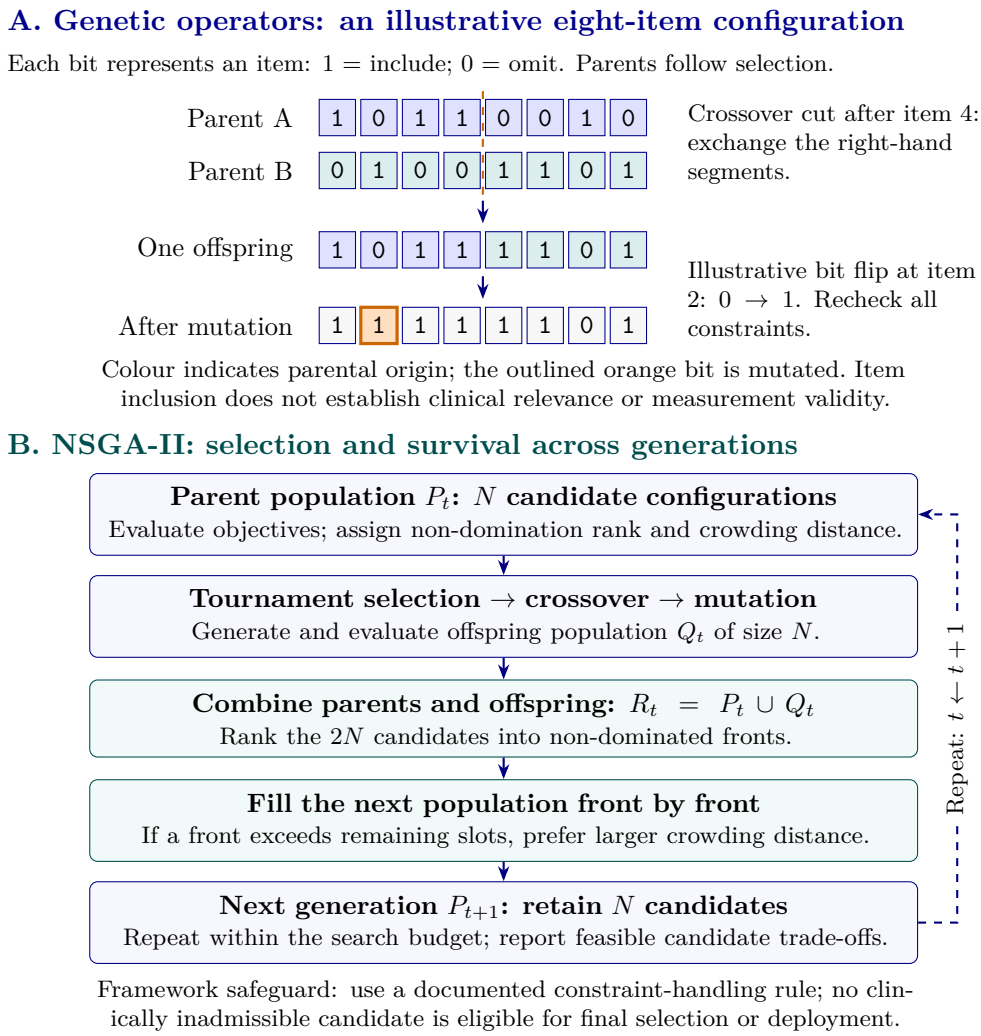



Figure 3: Genetic operators and NSGA-II survival, drawn for this article from the algorithmic principles described by Deb et al. [16]. Panel A shows an illustrative item-subset encoding, one of the two crossover offspring, and one possible mutation. Panel B shows elitist parent–offspring selection; crowding distance preserves diversity within a front. Here $t$ indexes generations and $N$ is population size. The bit strings are not patient data or optimised instruments, and no algorithm has been run to produce this illustration.

One proposed policy representation is

$$\mathbf{z}_i(t) = \pi_{\boldsymbol{\phi}}(\mathcal{H}_i(t), \mathbf{u}_i)\,, \tag{6}$$

where $\mathcal{H}_i(t)$ contains information available by time $t$, $\mathbf{u}_i$ records relevant patient preferences, and $\boldsymbol{\phi}$ parameterises a policy that may be optimised offline. Every resulting subset must satisfy the applicable content and measurement constraints. Approved common items or a supported calibrated metric should preserve comparability; raw sums from different subsets should not be compared as though they were the same score. Policy updates require versioning, patient choice, and evaluation before use. Changes to item composition must also be checked against the definitions of the compartmental burden states in Section 4.5.1.

A later extension could optimise $\Delta$ and $\eta$, balancing missed target events and false alerts against response burden and available review capacity. Such optimisation should operate only within prospectively approved schedules and routine-review purposes. It is not unconstrained tuning of emergency thresholds or automated treatment allocation. Historical data may not identify the consequences of alternative sampling frequencies or clinical responses; simulation assumptions and subsequent prospective testing must therefore be explicit. Longer intervals that make recall periods inappropriate, or thresholds that generate unactionable alerts, should be rejected regardless of their apparent numerical benefit. This keeps the contribution focused on the methodology for selecting inputs and configuring care, distinct from developing the underlying predictive model.

## 4.3 Design, administration, and measurement

The design record should distinguish faithful electronic migration from changes to instrument content. The 2023 ISPOR recommendations support using existing comparability evidence when it is sufficient and appropriate migration practices are followed; they do not require a new equivalence study for every electronic implementation [3]. Conversely, changes to wording, recall periods, response options, language, or scoring require a reasoned assessment of whether existing evidence remains applicable.

For rural implementation, the proposed design process includes usability work with intended users on available devices and under realistic connection conditions. It should test readability, navigation, saving interrupted responses, private completion, and whether reminders expose sensitive information. Alternative modes or assisted completion should be offered where feasible, with the mode and type of assistance recorded. Assistance that enables a patient to give their own response should be distinguished from a proxy answering on the patient's behalf. The latter is not silently relabelled as self-report.

Administration frequency should match both the construct's recall period and the service's ability to use the information. Any optimised or adaptive configuration from Section 4.2.1 remains subject to these measurement and usability requirements. Software should preserve instrument and scoring versions, acquisition time, and the time a response becomes available to clinicians. These are proposed safeguards for interpretation and traceability; their feasibility should be evaluated rather than presumed.

## 4.4 Data quality, missingness, and longitudinal change

Data review should start with the eligible population, not the final analytic file. The proposed audit follows eligibility, invitation, participation, completion, clinical review, and action, retaining

explicit denominators at each stage. Item non-response, a missed scheduled assessment, delayed synchronisation, and permanent withdrawal are different events. Where acceptable and feasible, reasons should be recorded without attributing non-completion to motivation or symptom change by default.

A related example from hospital-record research is the descriptive analysis of ICD-coded sepsis in MIMIC-IV by Okpara et al. [20]. The authors distinguish all eligible admissions from the first eligible admission per patient, specify in-hospital, 30-day, and 90-day mortality windows, and describe the availability of candidate physiological variables. For this framework, the methodological connection is the need to document the unit of analysis, eligibility rules, outcome timing, and data availability before modelling. This is an analogy across clinical settings, not evidence that sepsis mortality associations or hospital-record definitions transfer to rural mental-health ePROMs.

Missing-data methods depend on the question and assumptions. Multiple imputation can be useful under appropriate assumptions, but it does not remove the need to explain the missingness process or examine sensitivity to departures from those assumptions [21]. For longitudinal analyses, the plan should account for repeated observations, irregular timing, and relevant service-level clustering. For predictive modelling, processing learned from data must be fitted within the development or resampling partition; a prospective prediction must not use information unavailable at its prediction time [6, 7]. In this framework, imputed analytical values must remain distinguishable from observed responses and should not be displayed as statements made by the patient.

Interpretation of change should distinguish measurement error, a potentially important change, and evidence of a treatment effect. Reliable-change approaches assess whether change exceeds expected measurement noise under their assumptions; they do not by themselves establish that the change matters to the patient or was caused by care [22]. The proposed clinical display should therefore show actual observation times, gaps, score direction, and relevant contextual information. Where individual-level interpretation is unsupported, the uncertainty should remain visible rather than being converted into a definitive improvement or deterioration label.

Let $Y_{ij}$ denote the observed score for patient $i$ at assessment time $t_{ij}$, with $Y_{i0}$ the baseline score. Under the classical reliable-change assumptions, an illustrative calculation is [22]

$$\Delta Y_{ij} = Y_{ij} - Y_{i0}, \qquad \mathrm{RCI}_{ij} = \frac{\Delta Y_{ij}}{\sigma_{\mathrm{ref}}\sqrt{2(1 - r_{\mathrm{rel}})}}. \tag{7}$$

The reference-population score standard deviation is $\sigma_{\mathrm{ref}} > 0$, and $0 \leq r_{\mathrm{rel}} < 1$ is an appropriate reliability estimate, such as test–retest reliability in a relevant stable population. This form assumes equal measurement-error variances and independent errors at the two assessments. Under an approximately normal no-change distribution, $|\mathrm{RCI}_{ij}| > 1.96$ is a conventional two-sided 95% criterion for a single comparison. It is not a clinical alarm threshold, an adjustment for repeated testing, or proof of meaningful improvement. The sign must be interpreted using the instrument's scoring direction, and different error structures require an adapted denominator.

### 4.5 Disease modelling: trajectories, illness states, and progression

The disease-modelling decision begins with the question: which aspect of illness course must be represented to support the intended use? The proposed specification should identify the outcome or estimand, time origin, observation schedule, target population, care conditions, and whether the purpose is description, prognosis, or scenario exploration. In mental-health applications, the

representation should allow improvement and recurrence where relevant rather than impose a one-directional progression. A sequence of ePROM scores is a set of observations, not automatically a disease model.

Three complementary approaches can be considered, according to the question and available evidence. The equations below are candidate specifications, not models fitted to patient data. No empirical coefficients, transition rates, or clinical thresholds are reported; the later worked example uses hypothetical inputs:

1. **Longitudinal trajectory models.** Mixed-effects models account for repeated observations within individuals and can represent variation in trajectories [12]. For this framework, candidate uses include describing symptom burden or functioning over follow-up. The record should justify the time scale, functional form, and handling of within-person and service-level dependence. A modelled score trajectory must not be relabelled as a biological disease mechanism.
   A Gaussian mixed-effects specification is

$$\begin{aligned} Y_{ij} &= m_i(t_{ij}) + \varepsilon_{ij}, \\ m_i(t) &= \mathbf{x}_i(t)^\mathsf{T}\boldsymbol{\beta} + \mathbf{z}_i(t)^\mathsf{T}\mathbf{b}_i, \\ \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{8}$$

   Here, $m_i(t)$ is the underlying score trajectory, $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ are specified covariate vectors, $\boldsymbol{\beta}$ contains population-level coefficients, and $\mathbf{b}_i$ contains patient-specific deviations with covariance matrix $\mathbf{D}$. Residuals $\varepsilon_{ij}$ are assumed mutually independent and independent of $\mathbf{b}_i$ in this example. Serial dependence, service-level clustering, or ordinal and bounded outcomes may require different specifications. Including treatment as a covariate does not, by itself, identify its causal effect.

2. **Multi-state models.** These describe movement among specified states; hidden-state extensions can represent states observed through imperfect markers [13]. Candidate mental-health applications could examine sustained symptomatic periods and clinically established remission or recurrence. State definitions, permitted transitions, and any dependence on history or time already spent in a state require justification. An ePROM threshold alone does not establish clinical remission, relapse, or a newly discovered disease subtype.
   For a continuous-time Markov specification, let $S_i(t)$ be the patient's state at time $t$. With covariates suppressed, the instantaneous transition intensity from state $a$ to state $b$ is

$$q_{ab}(t) = \lim_{\delta \downarrow 0} \frac{\Pr\{S_i(t+\delta) = b \mid S_i(t) = a\}}{\delta}, \qquad a \neq b. \tag{9}$$

   When the intensity matrix $\mathbf{Q}$ is constant over the interval of interest,

$$q_{aa} = -\sum_{b \neq a} q_{ab}, \qquad \mathbf{P}(\tau) = \exp(\tau \mathbf{Q}), \quad \tau \geq 0. \tag{10}$$

   The exponential is a matrix exponential, and entry $P_{ab}(\tau)$ is the probability of occupying state $b$ after duration $\tau$, conditional on starting in state $a$. Intensities are rates, not probabilities. This formulation assumes that the current state sufficiently summarises the relevant history; time-varying intensities or dependence on duration in a state require extensions. Clinical state definitions remain a separate measurement requirement.

3. **Joint longitudinal and time-to-event models.** These connect repeated measurements with an event process and can update prognostic estimates as measurements accumulate [14]. A

potential application is relating symptom trajectories to a separately defined clinical event. The record should specify event ascertainment, censoring, and the assumed association between the processes. Joint modelling does not automatically resolve an informative visit schedule or missingness unrelated to the modelled event.

One current-value association links the trajectory in Equation (8) to an event hazard:

$$h_i(t \mid \mathbf{b}_i) = h_0(t) \exp\Big\{\boldsymbol{\gamma}^{\mathsf{T}} \mathbf{w}_i + \alpha m_i(t)\Big\}. \tag{11}$$

Here, $h_0(t)$ is the baseline hazard, $\mathbf{w}_i$ contains specified baseline covariates, $\boldsymbol{\gamma}$ their coefficients, and $\alpha$ an association parameter, not a causal effect. A dynamic prediction target is

$$p_i(t, u) = \Pr\{T_i \leq u \mid T_i > t, \mathcal{H}_i(t)\}, \qquad u > t, \tag{12}$$

where $T_i$ is the event time and $\mathcal{H}_i(t)$ contains observations and covariates available by prediction time $t$. The output is a conditional event probability, not the hazard itself. Estimation must account for latent-effect uncertainty and the specified censoring assumptions, without using future observations in an ostensibly prospective prediction.

For rural data, the proposed assessment should ask whether a recorded event represents illness course, service access, or both. A return visit or hospital admission should not be treated as an unqualified marker of deterioration without examining how travel, availability, and recording practices affect observation. Similarly, dropout is not recovery, delayed synchronisation is not a late symptom onset, and a gap between assessments does not reveal the exact time of a transition. Treatment changes and comorbidity should be recorded where relevant. Models fitted during ongoing care describe that setting; they do not automatically estimate untreated natural history or causal treatment effects.

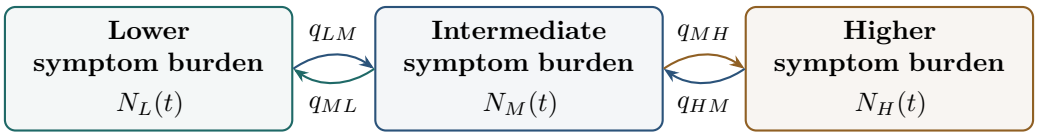
The required output is a disease-model specification and validation plan, or a documented decision to retain descriptive monitoring. The specification should identify data sources, structural assumptions, parameter uncertainty, and intended uses. Model-transparency guidance distinguishes clinical plausibility, code verification, and comparison with independent or prospectively observed data [23]. Here, checks should match the output: trajectory fit and prediction-interval coverage for scores; transition or event probabilities and their calibration for prognostic use; and sensitivity to state definitions, missingness, and model structure. Sparse events may justify simplifying or deferring a model rather than estimating unsupported local transition rates.

Disease-course simulations may subsequently explore service-planning scenarios, but their outputs should be labelled as conditional on the model and input evidence. Simulated reductions in relapse cannot establish intervention effectiveness. Any treatment-effect or policy claim requires a separate causal specification and appropriate evidence; predictive accuracy alone is insufficient. Uncertainty from a disease model must remain visible when its outputs are used as AI inputs or presented for clinical interpretation.

A related example from infectious-disease modelling is provided by Okpara et al. [24], who formulate a six-compartment Caputo fractional-order model and examine positivity, boundedness, stability, sensitivity, and illustrative numerical scenarios. Their numerical parameters are not estimated from a specific outbreak, so the simulations are mechanistic illustrations rather than empirical intervention-effect estimates. For the present framework, this example motivates attention to mathematical consistency and the distinction between simulation and validation; it does not justify transferring infectious-transmission mechanisms or reproduction-number thresholds to mental-health symptom burden. The model below remains an ordinary continuous-time state-transition model, not a fractional-order extension.

#### 4.5.1 Compartmental representation of symptom burden

A transparent starting point for the proposed analysis is a reversible, three-compartment cohort model (Figure 4). State-transition guidance supports defining health states and their connections around the decision question [25]. Here, the illustrative compartments are *lower* ($L$), *intermediate* ($M$), and *higher* ($H$) symptom burden. Their thresholds must be specified for the selected instrument, population, and purpose, with patient and clinical input; no numerical cut-offs are assumed. These are proposed burden categories, not diagnoses, remission criteria, or validated biological stages. If categorisation is unsupported, the continuous trajectory model should be retained instead.




$N_L(t) + N_M(t) + N_H(t) = N$ Rates $q_{ab}$ are measured per month.
Both worsening and improvement are permitted; no transmission mechanism is assumed.

Figure 4: Proposed three-compartment symptom-burden model for illustrating Domain 5. Compartments describe mutually exclusive states within a closed cohort, not validated diagnoses. Arrows denote instantaneous per-person transition rates. Only adjacent direct transitions are included in this starting structure; movement from $L$ to $H$ across an observation interval can still occur through $M$. State definitions and the need for additional transitions must be assessed before empirical use.

Let $N_L(t)$, $N_M(t)$, and $N_H(t)$ denote expected numbers of patients in the three states at time $t$, measured in months. For constant non-negative transition rates, the cohort balance equations are

$$\begin{aligned}
\frac{dN_L(t)}{dt} &= -q_{LM}N_L(t) + q_{ML}N_M(t), \\
\frac{dN_M(t)}{dt} &= q_{LM}N_L(t) - (q_{ML} + q_{MH})N_M(t) + q_{HM}N_H(t), \\
\frac{dN_H(t)}{dt} &= q_{MH}N_M(t) - q_{HM}N_H(t).
\end{aligned} \tag{13}$$

Each term is an expected patient flow per month. The derivatives sum to zero, preserving $N_L + N_M + N_H = N$. This is a closed-cohort approximation under the stated care conditions, not an epidemic model or an estimate of untreated natural history. Entry and mortality require explicit extensions when relevant. Missing assessments require an appropriate observation or missing-data model; they must not be interpreted as improvement.

Using the row-vector convention $\mathbf{n}(t) = (N_L(t), N_M(t), N_H(t))$, the corresponding generator and projection are

$$\begin{gathered}
\mathbf{Q} = \begin{pmatrix} -q_{LM} & q_{LM} & 0 \\ q_{ML} & -(q_{ML} + q_{MH}) & q_{MH} \\ 0 & q_{HM} & -q_{HM} \end{pmatrix}, \\
\frac{d\mathbf{n}(t)}{dt} = \mathbf{n}(t)\mathbf{Q}, \qquad \mathbf{n}(t+\tau) = \mathbf{n}(t)\exp(\tau\mathbf{Q}).
\end{gathered} \tag{14}$$

This is the cohort form of Equation (10). Zero direct $L$–$H$ intensities do not imply zero finite-interval $L$–$H$ transition probabilities. The proposed analysis should compare this parsimonious structure with clinically justified alternatives, rather than fix its structural zeros solely for convenience.

#### 4.5.2 Planned calculations and a hypothetical worked example

The proposed calculations will begin with initial state counts, observation times, transition information, and the period over which rates can reasonably be treated as constant. With completely observed transition paths, a basic constant-rate estimate is

$$\widehat{q}_{ab} = \frac{D_{ab}}{\mathrm{PT}_a}, \qquad a \neq b, \tag{15}$$

where $D_{ab}$ counts transitions from $a$ to $b$ and $\mathrm{PT}_a$ is total observed patient-time in the origin state. The denominator is state-specific exposure, not the number of questionnaires or the total cohort size.

Repeated ePROMs usually provide states at visits rather than exact transition paths. For that situation, the proposed calculation uses the panel-data likelihood instead of counting observed changes as all transitions [13]:

$$\ell(\mathbf{Q}) = \sum_i \sum_{j=1}^{J_i} \log\Big( [\exp\{(t_{ij} - t_{i,j-1})\mathbf{Q}\}]_{s_{i,j-1}, s_{ij}} \Big). \tag{16}$$

Here, $s_{ij}$ is the observed state at visit $j$, and $J_i$ is the number of intervals after baseline for patient $i$. Maximisation is conditional on initial states, the specified transition structure, and appropriate observation assumptions. This likelihood allows unobserved intermediate transitions between visits. Misclassified states, informative attendance, or relevant clustering require further specification; last-observation-carried-forward exposure is not an automatic solution.

**Hypothetical arithmetic only.** Table 1 supplies invented counts and exposures solely to demonstrate Equation (15). They are not observations, literature estimates, or proposed clinical benchmarks. Assume independently, for illustration, an initial cohort of 120 patients with $\mathbf{n}(0) = (60, 40, 20)$.

Table 1: Entirely hypothetical complete-path inputs for basic rate calculations. The middle-state exposure is the same denominator for both outgoing middle-state transitions; it is not counted twice when calculating total follow-up.

| Transition | Illustrative count | Origin-state patient-months | Rate calculation (month$^{-1}$) |
|---|---|---|---|
| $L \to M$ | 12 | 300 | $q_{LM} = 12/300 = 0.04$ |
| $M \to L$ | 20 | 200 | $q_{ML} = 20/200 = 0.10$ |
| $M \to H$ | 6 | 200 | $q_{MH} = 6/200 = 0.03$ |
| $H \to M$ | 8 | 100 | $q_{HM} = 8/100 = 0.08$ |

At baseline, the expected flows are $0.04 \times 60 = 2.4$ from $L$ to $M$, $0.10 \times 40 = 4.0$ from $M$ to $L$, $0.03 \times 40 = 1.2$ from $M$ to $H$, and $0.08 \times 20 = 1.6$ from $H$ to $M$, all in patients per month. A first-order Euler approximation illustrates the arithmetic:

$$\mathbf{n}(t + \Delta t) \approx \mathbf{n}(t) + \Delta t\, \mathbf{n}(t)\mathbf{Q}. \tag{17}$$

For $\Delta t = 1$ month, this gives

$$\begin{aligned} \mathbf{n}(0)\mathbf{Q} &= (1.6, -1.2, -0.4), \\ \mathbf{n}_{\mathrm{Euler}}(1) &= (60, 40, 20) + (1.6, -1.2, -0.4) \\ &= (61.6, 38.8, 19.6). \end{aligned} \tag{18}$$

Euler calculations are explanatory, not the preferred estimator or projection method. Large time steps can produce invalid negative counts. For constant rates, the matrix exponential accounts for changes in occupancy and possible repeated transitions within the interval:

$$\mathbf{n}(1) = (60, 40, 20)\exp(\mathbf{Q}) \approx (61.512, 38.889, 19.599). \tag{19}$$

In Equation (19), the interval is one month and $\mathbf{Q}$ is expressed per month. The entries are expected counts, not observed or fractional individual patients; they sum to 120, apart from any rounding.

The analysis will also calculate state proportions and cumulative time in the higher-burden state:

$$\pi_H(t) = \frac{N_H(t)}{N}, \qquad \mathrm{PM}_H(T) = \int_0^T N_H(t)\,dt, \tag{20}$$

where $T$ is the projection horizon and $\mathrm{PM}_H$ is measured in patient-months. With the same hypothetical inputs, $\pi_H(1) \approx 0.1633$, and the model projects approximately 112.90 higher-burden patient-months over six months. These are conditional model outputs, not evidence of treatment effectiveness, numbers of unique patients, or required clinical visits.

Before empirical interpretation, the proposed analysis will check conservation, non-negative occupancies, and agreement between the balance equations and matrix projections; assess sensitivity to rate and state definitions; propagate parameter uncertainty with resampling appropriate to the study design; and compare projected state occupancy with independent follow-up where feasible. No uncertainty intervals are supplied for the invented example, because it contains no empirical information about parameter precision. This calculation sequence operationalises the disease-modelling gate without claiming that the compartment structure has been validated.

### 4.6 AI suitability and comparative value

AI consideration begins with a decision problem, not a preferred algorithm. Optimising the ePROM inputs or monitoring configuration is a separate task from estimating the predictive model; the complete combined procedure must nevertheless be evaluated without data leakage. Where disease modelling is justified, the comparison should establish whether AI improves a specified trajectory, transition, or event prediction over an appropriate statistical model using the same information. Where direct prediction is sufficient, a separate disease model need not be imposed. The proposed record should define the prediction target, horizon, eligible population, available predictors, intended action, and comparator. A model predicting a future ePROM score addresses a different question from one predicting an independently assessed clinical event. If the target is an ePROM threshold, the resulting claim must remain about that threshold rather than being relabelled as diagnosis, recovery, or crisis prevention.

The framework requires comparison with a credible simpler approach, such as existing monitoring, a transparent change rule, or a conventional statistical model appropriate to the same prediction target. Model-study reporting and risk-of-bias assessment should use TRIPOD+AI and PROBAST+AI, respectively [6, 7]. The proposed evaluation should address calibration and discrimination, uncertainty, relevant subgroup performance, and applicability to the intended rural population. Partitioning must reflect the deployment question, protecting against patient overlap and temporal leakage; evaluation across sites or later periods should be planned when those are the intended forms of generalisation.

For a fully observed binary event $E_i \in \{0, 1\}$ over a prespecified horizon, let $\widehat{p}_i$ be the predicted event probability and $n$ the number of validation patients. Overall probability error can be summarised by the Brier score [26]:

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{p}_i - E_i)^2. \tag{21}$$

Lower values indicate smaller squared probability errors within the same evaluation population and prediction task. This measure is not a test of clinical usefulness. For $0 < \widehat{p}_i < 1$, a logistic calibration assessment is

$$\mathrm{logit}\{\Pr(E_i = 1 \mid \widehat{p}_i)\} = a_{\mathrm{cal}} + b_{\mathrm{cal}} \,\mathrm{logit}(\widehat{p}_i), \tag{22}$$

where $\mathrm{logit}(p) = \log\{p/(1-p)\}$. The benchmark is $a_{\mathrm{cal}} = 0$ and $b_{\mathrm{cal}} = 1$, but a calibration curve is still needed to detect departures hidden by this two-parameter summary. These assessments require held-out or otherwise appropriately validated predictions, not just apparent fit. In time-to-event applications with censoring, the simple complete-outcome Brier calculation requires a justified censoring-adjusted counterpart; unobserved events must not be coded as absent.

Decision-curve analysis can compare model-guided actions over clinically meaningful threshold preferences, but estimated net benefit is not a substitute for observing effects in a deployed service [27]. Here, review workload, unavailable follow-up capacity, and the consequences of false reassurance or unnecessary escalation are additional decision criteria. A technically more accurate model should not progress when these consequences make its proposed use unacceptable. Sparse local evidence should lead to narrower claims, further evaluation, or non-adoption—not an assumption that performance transfers. Prognostic accuracy alone is insufficient to justify personalised treatment recommendations; that would require a separate causal question and appropriate evidence.

For the same complete-outcome binary setting, net benefit at a clinically justified threshold $p_{\mathrm{th}} \in (0, 1)$ can be written as [27]

$$\mathrm{NB}(p_{\mathrm{th}}) = \frac{\mathrm{TP}(p_{\mathrm{th}})}{n} - \frac{\mathrm{FP}(p_{\mathrm{th}})}{n} \frac{p_{\mathrm{th}}}{1 - p_{\mathrm{th}}}. \tag{23}$$

Here, TP and FP count patients whose predictions exceed the action threshold and who do or do not experience the target event, respectively. The threshold odds express the relative penalty for an unnecessary action compared with the benefit of an appropriate one. Compare the proposed model with relevant existing strategies and, where applicable, acting on everyone or no one over a justified threshold range. A positive value alone does not establish superiority, affordability, or effectiveness of the downstream intervention.

### 4.7 Interpretation, uncertainty, and explainability

The interpretation layer should keep four elements distinct: the patient's response, the derived score, any model output, and the proposed clinical response. Disease-model outputs should identify the trajectory, state, or event being estimated and distinguish observations from inferred states and forecasts. A clinician and patient should be able to see which is which. For a prediction, the display should identify the target and time horizon, relevant limitations, and the extent of supporting evidence. An explanation of contributing variables should not be presented as proof of causation or correctness. Published methodological criticism of current explainability approaches cautions against assuming that a plausible explanation establishes clinical trustworthiness [28].

The proposed usability evaluation should test whether users can recognise missing or stale information, question a prediction, and avoid over-reliance on an apparently precise number. Patient-facing explanations should be understandable without requiring familiarity with model terminology. Where data are insufficient or outside the evaluated scope, the system should be able to withhold a prediction while preserving an ordinary clinical review route. Such abstention is a proposed safety mechanism whose frequency and distribution require monitoring; it must not become an unnoticed way of excluding harder-to-measure patients.

### 4.8 Clinical decision-making and response

Every actionable output should have a named recipient, a defined review interval, an escalation route, and a documented way to communicate the response to the patient. Where alert thresholds or monitoring intervals are optimised, the permitted search space must reflect these responsibilities and response-capacity limits rather than maximise sensitivity alone. The clinical team should distinguish an unreviewed notification from a reviewed assessment and a completed action. Decisions should remain open to clinical context and patient preference; a score or prediction should not automatically determine diagnosis, treatment, or discharge.

The framework proposes that services rehearse missed notifications, staff absence, delayed responses, and system outages before introducing the relevant function. Collection of information that may indicate urgent distress requires a locally approved response plan and clear communication about monitoring hours and limitations. This is not an emergency-care protocol, and no generic ePROM or AI threshold is proposed for that purpose. Ethical oversight and early clinical evaluation should attend to accountability and human interaction, not only software performance [8, 29]. The principal progression question is whether an appropriate response can actually be delivered, not whether a dashboard successfully generates alerts.

### 4.9 Implementation and continuing governance

Implementation should be treated as continuing service work. NASSS draws attention to non-adoption, abandonment, and sustainability of health technologies, while the updated CFIR supports assessment of contextual determinants [10, 11]. The proposed local implementation plan should identify staffing, training, workflow integration, interoperability, technical support, recurrent costs, and responsibility when a supplier or system changes. Documentation effort should be proportionate: a low-risk display of established scores and a model directing scarce clinical attention should not receive identical scrutiny.

The governance record should specify authorised data uses, access, retention, secondary-use arrangements, version control, and incident response, subject to the relevant local requirements. WHO guidance provides an ethical basis for attention to autonomy, accountability, and inclusiveness [29]. Where Indigenous peoples or data are involved, governance must be developed with the relevant peoples and authorities rather than reduced to a generic consultation exercise. The CARE Principles foreground collective benefit, authority to control, responsibility, and ethics [30]. Their relevance is not limited to rural communities, and not every rural project involves Indigenous data.

Model or instrument updates should trigger a documented assessment of which earlier evidence remains applicable. The proposed change process includes testing before release, identifiable versions,

the ability to revert, and communication to affected users. Ongoing monitoring and withdrawal plans are consistent with lifecycle-oriented AI guidance [9].

### 4.10 Evaluation, equity, and revision

Evaluation must distinguish framework quality, implementation outcomes, and effects on care. Acceptability and adoption are not equivalent to clinical effectiveness, as implementation-outcome scholarship makes explicit [31]. Section 6 proposes separate studies for these questions. Successful early use would support a claim of feasibility within the observed setting, not a general claim of improved mental health.

Equity is assessed across the pathway. STANDING Together addresses how datasets represent people and how their limitations are documented and considered during use [32]. Building on that concern, this framework proposes examining who is invited, who can respond, whose data are usable, who receives a prediction, and whose resulting need receives a response. Relevant groups should be agreed locally, with attention to privacy and the precision available from small samples. A non-significant difference is not evidence of equivalence, and pooled performance should not obscure uncertainty for a small group. Conversely, publicly reporting finely divided rural cells can compromise privacy. The evaluation plan should specify both protected analyses and an appropriate disclosure approach.

For an agreed reporting period, a simple participation indicator and between-group difference are

$$\mathrm{Reach}_g = \frac{N_g^{\mathrm{complete}}}{N_g^{\mathrm{eligible}}}, \qquad \mathrm{Gap}_{g,g_0} = \mathrm{Reach}_g - \mathrm{Reach}_{g_0}. \tag{24}$$

The numerator counts eligible people in group $g$ who complete at least one scheduled assessment, not the total number of questionnaire submissions; $g_0$ is a prespecified reference group. The ratio is undefined when the eligibility denominator is zero. The gap is a descriptive participation difference, not a causal effect or a complete equity assessment. Report denominators and uncertainty, and separately examine sustained participation, review, and action. Similar model accuracy among respondents cannot compensate for unequal access to participation.

## 5 Illustrative application: deciding not to add AI

Consider a hypothetical adult rural outpatient service that wants to use repeated ePROMs to support planned reviews for people receiving care for depression or anxiety. Assume variable connectivity, shared devices for some patients, a small clinical team, and an externally developed prediction model offered by a supplier. These assumptions describe a constructed example only; no service, participants, implementation, or outcome data are represented.

Applying the framework would first narrow the purpose to helping patients and clinicians discuss change at scheduled reviews. The service would select instruments only after examining their construct coverage, population evidence, and burden. It would compare electronic and alternative completion routes, record assistance and transmission delays, and decide what staff can realistically review. It would avoid promising real-time surveillance merely because repeated data can be collected.

Table 2: Proposed progression gates. These are decision prompts, not validated cut-offs or a scored accreditation tool.

| Gate | Evidence needed for the intended use | Revise or stop the component when |
|---|---|---|
| Purpose and access | A specified clinical decision, feasible response, and participation routes for intended users. | No actionable purpose exists, or unresolved exclusion makes the proposed service inappropriate. |
| Measurement | Justified construct, instrument, administration mode, and interpretation; uncertainties recorded. | The intended interpretation exceeds the measurement evidence, with no acceptable plan to address the gap. |
| Data and analysis | Defined denominators, timestamps, missingness plan, and an analysis consistent with information available at use. | Data errors, leakage, or unsupported assumptions make outputs misleading. |
| Disease modelling | Defined trajectories, states or events; supported assumptions; fit-for-purpose validation and uncertainty assessment. | Clinical labels lack supporting evidence, event data are inadequate, or the model cannot support its intended use. Retain descriptive monitoring where appropriate. |
| Optimisation | Prespecified objectives, valid candidate configurations, protected constraints, and independent comparison with simpler selection. | Search gains depend on invalid shortened measures, data leakage, unacceptable workload, or unresolved subgroup harms. |
| AI added value | Comparison with a credible simpler option; applicable performance evidence and consequence assessment. | Incremental value is not supported, or expected harms and workload are unacceptable. Retain a non-AI route where suitable. |
| Clinical readiness | Named reviewers, rehearsed response routes, understandable outputs, and a safe fallback. | Alerts cannot be reviewed as specified or the required response is unavailable. |
| Sustainability | Agreed ownership, resources, monitoring, change control, and withdrawal arrangements. | Critical safety, governance, or performance concerns remain unresolved. |

At the disease-modelling gate, the service would decide whether it needs only a display of observed change, a model of symptom trajectories, or estimates of transitions between clinical states. If the available data contain ePROM scores but no defensible ascertainment of remission or recurrence, it should not fit and label a remission–relapse model on that basis. A future longitudinal analysis could address score trajectories if its assumptions and data requirements are met; this would remain distinct from diagnosing illness states or estimating treatment effects. No such analysis is performed in this example.

Suppose the hypothetical supplier provides overall model discrimination but no relevant local calibration, no comparison with an existing non-AI review process, and insufficient information on patients who do not complete questionnaires. Under the proposed AI gate, these omissions would prevent a justified transition to clinical use of that model. A reasonable recorded decision would be to defer AI and consider a carefully scoped, non-AI ePROM pathway, itself subject to the measurement and clinical-readiness gates. Future research could evaluate the model using appropriate permissions and a prospective design without exposing its outputs to clinical decision-makers initially.

The resulting decision record would identify the unavailable evidence, the person responsible for seeking it, and the conditions under which the decision would be revisited. An offline-first interface, for example, would not be judged solely by completion rates: delayed availability could make a score unsuitable for a time-sensitive prediction even when the patient completed it on time. This illustrates a dependency across design, data quality, and clinical action. It demonstrates the logic of the framework, not that the proposed pathway is safe, feasible, effective, or acceptable in practice.

# 6 A staged strategy for evaluating the framework

Three evaluands must remain separate: the *framework* as a decision aid; a particular *ePROM-based system*, including any disease model or AI component, built or selected using it; and the *care intervention* through which that system is used. The validity of a disease-course representation and the accuracy of any predictions are separate questions within system evaluation. A strong model does not validate the framework, and a well-received framework does not establish a clinical benefit. The following programme is proposed for future research; none of its phases has been completed for this article.

## 6.1 Phase 1: participatory refinement and content appraisal

Recruit a purposively diverse group of rural service users, carers where appropriate, clinicians, community representatives, measurement specialists, implementation researchers, and AI or data-governance specialists. Include people who do not use digital health services and provide non-digital participation routes. Ask participants to identify omissions, ambiguous terms, unrealistic responsibilities, and potential harms in the framework and decision record.

Use cognitive walkthroughs and structured discussions to assess relevance, comprehensibility, completeness, and feasibility. A subsequent consensus exercise could prioritise requirements, but its composition, procedures, handling of disagreement, and decision rules should be prospectively specified. Consensus would support acceptability of the recommendations within that group, not prove their effectiveness. Produce a versioned revision log showing which suggestions were accepted, rejected, or remain unresolved, and why. Substantial unresolved objections from affected

communities should not disappear into an aggregate agreement score.

### 6.2 Phase 2: comparative testing of decision quality

Test the framework on standardised service-planning cases that vary connectivity, measurement evidence, disease-state definitions, event ascertainment, model performance, staffing, and data governance. Include cases in which an apparently accurate short form omits essential content or an optimised alert policy exceeds clinical capacity. Compare teams using the framework with teams given the underlying guidance documents without the integrated decision tool. This active comparator is important: comparing only with no guidance would not establish whether the integration adds value. Randomise teams where feasible, or otherwise account for order, prior expertise, and learning effects.

Develop an assessment rubric before testing, with input independent of the framework's developers. The proposed primary outcome is the proportion of critical, case-specific methodological or safety omissions identified and appropriately addressed. Secondary outcomes include unsupported progression decisions, quality of justification, completion time, decision confidence, and user burden. Use independent, masked assessment of de-identified decision records where feasible. Some cases may permit several reasonable decisions; assess the evidence and reasoning rather than agreement with a single developer-preferred answer. This phase tests whether the framework improves planning, not whether patients benefit.

### 6.3 Phase 3: prospective feasibility in rural services

Apply the refined framework in a small, deliberately varied set of services. The number of sites and participants should be justified by the uncertainties to be resolved, feasibility objectives, and precision required, rather than an arbitrary doctoral recruitment target. Collect decision records, observation or interview data, staff time, implementation costs, participation denominators, and incident reports. Examine whether the framework changes choices, identifies issues missed by existing processes, or adds disproportionate administrative burden.

For a selected disease model, prespecify assessment of its structure, code, and application-specific validity, with independent data where feasible [23]. For predictive use, evaluate trajectory forecasts or event probabilities over the intended horizon without using later observations as if they were available at the prediction time. When evidence is insufficient, record the unresolved question rather than reporting the framework itself as validated.

Where optimisation is used, compare the entire selection procedure with approved short forms and simpler search strategies before testing a resulting care pathway. Prespecify evaluation budgets, independent data, measurement checks, patient burden, subgroup performance, and workload. Neither a favourable Pareto plot nor successful algorithm convergence establishes clinical benefit.

If an AI component is being considered, distinguish prospective *silent* evaluation, in which outputs do not guide care, from early *live* use. Silent evaluation may assess data flows and predictive performance but cannot establish clinician behaviour or patient benefit. Early live evaluation requires the relevant approvals, safety arrangements, and stage-appropriate reporting; DECIDE-AI specifically concerns this live evaluation stage [8]. Prespecify progression and pause criteria with local governance, including unacceptable unreviewed outputs, serious privacy incidents, or unresolved

disparities. A feasibility study should not claim to rule out rare harms.

### 6.4 Phase 4: effectiveness, implementation, and economic evaluation

Two causal questions need different comparisons. To evaluate the added value of the framework, compare service planning and implementation supported by the framework with an appropriately resourced existing planning process. To evaluate a particular care intervention, compare the resulting ePROM-based pathway, with or without AI, with a clearly described alternative. A positive result in the second comparison cannot automatically be attributed to the framework because staffing, training, and other components may also differ.

A pragmatic cluster-randomised design may be suitable when contamination between clinicians or teams is likely and sufficient clusters are available. Where randomisation is infeasible, a controlled longitudinal design may be considered, with explicit limitations concerning confounding and secular change. Design selection should follow the programme theory, context, and principal uncertainty [17]. A later trial would require prospective registration, a protocol and analysis plan, justified sample size, and relevant trial reporting, including CONSORT 2025 and the AI-specific extension when applicable [33, 34].

The primary outcome and time horizon should be agreed with patients and services before recruitment. Candidate clinical outcomes include functioning, quality of life, or symptom burden; candidate service outcomes include timely review and continuity. Independent or consistently administered outcome collection should reduce the risk that changing completion behaviour alone appears to improve care. The analysis should address clustering, repeated measures, differential attrition, and prespecified subgroup questions. Clinical effects, implementation outcomes, and resource consequences should be reported separately. Economic analysis should count recurrent staffing, digital support, licensing, and costs borne by patients, not just the purchase price of a platform.

## 7 Discussion

The principal argument is that ePROMs and AI should be evaluated as parts of an accountable care pathway, not as a sequence in which technical success automatically licenses the next step. The proposed framework contributes a decision record, explicit progression gates, and a distinction between three evaluands. Disease modelling adds an explicit representation of illness course between measurement and analysis-method selection. Its value depends on whether that representation serves the clinical question, not on adding complexity or naming score categories as diseases. Its most consequential feature is that declining AI, changing an instrument, or reducing an unactionable monitoring burden can be legitimate outcomes of methodological work. Innovation is not equated with increasing algorithmic complexity.

Evolutionary algorithms are therefore positioned as optional methodological tools, not the centre of the research programme or a reason to rename its central contribution. The proposed innovation is how measurement evidence, competing objectives, admissible configurations, and human decisions are connected. Optimising inputs and implementation settings can complement a separate quantitative programme developing the predictor without duplicating its model-development objective. No claim is made that evolutionary methods are necessary or superior for every rural service.

Table 3: Proposed evaluation domains and the limits of their interpretation. Measures and thresholds require local specification and testing.

| Domain | Candidate indicator and denominator | Claim not established by this indicator alone |
|---|---|---|
| Framework utility | Critical issues appropriately addressed per assessed planning case; time per completed decision record. | Improved patient outcomes or validity of a particular instrument. |
| Patient-centredness | Patient-reported involvement in interpretation and care decisions; burden among participants and reasons for declining. | Acceptability to eligible people who were not reached or did not respond. |
| Reach and equity | Invitation, participation, completion, review, and action among the relevant eligible groups, with uncertainty. | Equity based only on average completion or on non-significant group differences. |
| Measurement | Evidence supporting the intended score and change interpretations in the relevant population and mode. | Clinical benefit from collecting or displaying the score. |
| Disease modelling | Fit of trajectories; plausibility of states and transitions; independent assessment of forecasts or event probabilities; structural sensitivity. | True disease states, untreated natural history, or causal treatment effects simply because the model fits observed data. |
| Optimisation | Independently evaluated loss–burden trade-offs, constraint satisfaction, search stability, and comparison with simpler methods. | A globally optimal subset, a validated shortened questionnaire, or better care from search fitness alone. |
| AI performance | Calibration, discrimination, prediction coverage, and clinical consequence estimates for the specified target. | Treatment benefit, causal explanations, or transfer to unassessed settings. |
| Safety and response | Unreviewed actionable outputs, response delays, incidents, and reasons for override per relevant encounter or output. | Absence of rare harm from a small feasibility sample. |
| Implementation | Acceptability, feasibility, use, costs, and sustained operation over a specified period. | Clinical effectiveness simply because the technology is adopted. |
| Care outcomes | Prespecified between-group differences in patient-important outcomes, with uncertainty and resource consequences. | Benefit attributable specifically to the framework unless the design isolates that comparison. |

The framework does not replace existing guidance. In particular, FUTURE-AI already addresses broad lifecycle questions, and NASSS and CFIR already foreground implementation context [9–11]. The intended additional value is narrower: linking patient-reported measurement decisions to longitudinal analysis and feasible clinical responses in a specified rural service. Whether this integration provides an advantage over using those resources directly is an empirical question addressed by the proposed active-comparator evaluation. No claim is made that this is the first framework to connect digital health, measurement, or equity.

A rural focus should change decisions rather than merely change the description of the setting. In the proposed approach, connection failures alter the interpretation of missing observations; limited response capacity changes whether a monitoring function is appropriate; language and cultural context influence instrument evidence requirements; and small populations affect both statistical precision and disclosure risk. These are conditional considerations to investigate locally. They should not be interpreted as fixed deficits of rural communities or as reasons to withhold useful services while waiting for unattainable evidence.

There is also a risk of over-engineering. Extensive documentation can consume resources that a small team needs for direct care. The framework therefore requires proportionate application, use of existing evidence where appropriate, and explicit measurement of its own burden. Shared methodological support across services may be more feasible than expecting each clinic to conduct independent psychometric and AI research. Nevertheless, resource constraints do not justify disguising uncertainty or allowing a nominally human-reviewed system to operate without meaningful review.

### 7.1 Limitations and research priorities

The framework is provisional. Its literature base is purposive and not systematically appraised; selection may overrepresent accessible English-language guidance and high-income settings. No patient or community partners participated in developing this draft, so its language and priorities cannot be described as co-designed. The illustrative example is constructed, and neither the gates nor the decision-record template has been tested. The initial adult outpatient scope limits direct transfer to children, acute care, and other clinical populations. The disease-modelling component identifies candidate methods rather than a validated mental-health natural-history model; the cited methodological examples do not establish their applicability to a particular rural population. The evolutionary-optimisation component has not been benchmarked, and no optimised short form, adaptive policy, or monitoring schedule has been validated. Questions specific to generative AI are not resolved by this framework.

A framework can improve documentation without improving decisions, and clearer decisions may still fail to improve outcomes when effective care is unavailable. These possibilities make independent evaluation essential. Priority research should establish whether the framework detects consequential omissions, whether intended users can apply it without excessive burden, and whether it improves choices compared with access to established guidance alone. Only after those questions are addressed should clinical or equity improvement be attributed to its use.

# 8 Conclusion

This article proposes a methodological framework connecting rural context, ePROM selection and design, longitudinal data interpretation, disease modelling, optional AI, clinical action, implementation, and evaluation. Its intended contribution is to make the justification and consequences of each transition visible. Measurement evidence, patient priorities, equitable participation, and the capacity to respond are prerequisites for a defensible pathway, not accessories to an algorithm. An optional multi-objective optimisation layer can support ePROM selection, personalisation, and monitoring configuration, but cannot replace measurement validation or clinical judgement. The framework offers a structured basis for participatory refinement and future testing; it does not yet establish improved care or outcomes.

# Declarations

**Ethics and consent.** This conceptual article reports no participant recruitment, individual-level data analysis, or clinical intervention. Future studies described here would require the relevant institutional review, governance arrangements, and consent processes before commencement. No ethics approval or exemption determination is claimed for this draft.

Declaration of AI Use

This manuscript was prepared through the intellectual contributions of the author(s) to the study design, data analysis, interpretation of results, and scholarly content. Grammarly and ChatGPT were used solely to assist with grammatical refinement and reference formatting during manuscript revision. The author(s) accept full responsibility for the accuracy, integrity, and final content of the manuscript.

APPENDIX

# A Source-to-domain crosswalk

The crosswalk identifies selected foundations for the proposed integration. It is not a systematic evidence map, a formal quality assessment, or a claim of endorsement by the developers of the cited resources.

Table 4: Selected methodological foundations and their roles in the proposed framework.

| Resource | Role of the original resource | Proposed use and boundary in this article |
|---|---|---|
| COSMIN [5] | Appraisal and synthesis of evidence about PROM measurement properties. | Supports selection and the specification of measurement uncertainties; not evidence that a named instrument is appropriate locally. |
| ISPOR mode-comparability guidance [3] | Assessment of when existing evidence supports comparability across collection modes. | Separates faithful migration from substantive modification; not a blanket requirement for new equivalence testing. |
| Longitudinal, multi-state, and joint models [12–14] | Methods for repeated measurements, state transitions, and linked longitudinal/event processes. | Informs choice of disease-course representation; does not establish clinical states or treatment effects from ePROM scores alone. |
| Model transparency and validation [23] | Documentation and application-specific assessment of healthcare models. | Informs the disease-model specification and validation plan; sensitivity analysis is not independent evidence of accuracy. |
| TRIPOD+AI [6] | Reporting of clinical prediction-model studies using regression or machine learning. | Makes model evidence inspectable; checklist completion is not a deployment authorisation. |
| PROBAST+AI [7] | Quality, risk-of-bias, and applicability assessment for prediction-model development and evaluation. | Structures appraisal of candidate evidence; does not establish effectiveness of a clinical workflow. |
| DECIDE-AI [8] | Reporting of early live clinical evaluation of AI decision support. | Supports transparency during limited clinical use; not a label for silent evaluation or a substitute for safety review. |
| FUTURE-AI [9] | Consensus guidance for trustworthy and deployable healthcare AI. | Informs lifecycle questions and monitoring; the present framework adds a proposed ePROM-to-action decision record. |
| STANDING Together [32] | Consensus recommendations on health-dataset documentation and use, including diversity and bias. | Informs representation and dataset limitations; the proposed audit extends to participation, review, and action. |
| NASSS and updated CFIR [10, 11] | Theorisation of technology sustainability and assessment of implementation determinants. | Informs contextual and organisational assessment; does not validate measurement or prediction. |
| WHO and CARE [29, 30] | Ethical AI governance and Indigenous data-governance principles, respectively. | Informs accountability and appropriate authority; does not replace jurisdiction-specific or community-specific requirements. |
| Complex-intervention guidance [17] | Development and evaluation attentive to context, programme theory, uncertainty, and resources. | Informs the staged research agenda; no effectiveness inference follows from conceptual alignment. |
| Implementation outcomes [31] | Distinguishes implementation outcomes from service and clinical outcomes. | Supports separate evaluation claims; adoption is not treated as patient benefit. |

## B Reusable decision-record template

Complete one record for each consequential choice and update it when assumptions, instruments, software, evidence, or service arrangements change. The template is a proposed working aid, not a validated instrument.

1. **Purpose and scope:** What clinical or methodological decision is being supported? For whom, where, by whom, and over what time horizon? Which uses are excluded?
2. **Alternatives:** What existing, non-digital, non-AI, or simpler approach could meet the same purpose? Why is change being considered?
3. **Optimisation, if used:** What is being selected or configured? Specify objectives, clinical and measurement constraints, permitted schedules, comparators, search settings, and the independent evaluation plan. Record who chooses among feasible trade-offs; do not equate search fitness with clinical benefit.
4. **Evidence:** Which sources support the choice? Which are directly applicable, indirect, conflicting, or unavailable? What remains uncertain?
5. **Measurement and data:** Which construct, instrument version, language, mode, schedule, and scoring method apply? How are timestamps, missingness, and assistance recorded?
6. **Disease modelling:** Is a formal model needed? Which trajectories, states, events, time origin, and care conditions apply? What supports their clinical interpretation? Record assumptions, event ascertainment, validation, uncertainty, and whether a simpler representation is sufficient.
7. **Participation and equity:** Who may be excluded at invitation, completion, analysis, review, or response? What alternative route and mitigation are available?
8. **Interpretation and response:** What will the output mean? Who reviews it, within what interval, and with what available action? What happens if information is stale or insufficient?
9. **Governance and resources:** Who authorises use, controls access and change, and funds continuing work? Which relevant community and patient perspectives informed the choice?
10. **Decision and conditions:** Proceed within scope, revise or gather evidence, or do not implement this component. Record the rationale, unresolved objections, and approvals required.
11. **Evaluation and review:** Which outcomes, denominators, and harms will be monitored? Who owns the decision? What review date, incident, or change triggers reassessment?

## References


[1] Khushi Kohli, Bhav Jain, Tej A. Patel, Hatice Nur Eken, Edward Christopher Dee, and John Torous. The digital divide in access to broadband internet and mental healthcare. *Nature Mental Health*, 2:88–95, 2024. doi: 10.1038/s44220-023-00176-z.

[2] Olivia J. Fisher, Kelly McGrath, Caroline Grogan, Wendell Cockshaw, and Chez Leggatt-Cook. Care navigation addresses issues of tele-mental health acceptability and uptake in rural and remote Australian communities. *PLOS ONE*, 19(4):e0298655, 2024. doi: 10.1371/journal.pone.0298655.

[3] Paul O'Donohoe, David S. Reasner, Sarrit M. Kovacs, Bill Byrom, Sonya Eremenco, Alexandra I. Barsdorf, Valdo Arnera, and Stephen Joel Coons. Updated recommendations on evidence needed to support measurement comparability among modes of data collection for patient-reported outcome measures: A good practices report of an ISPOR task force. *Value in Health*, 26(5):623–633, 2023. doi: 10.1016/j.jval.2023.01.001.

[4] Jaime Delgadillo, Kim de Jong, Mike Lucock, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: A multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry*, 5(7):564–572, 2018. doi: 10.1016/S2215-0366(18)30162-7.

[5] Lidwine B. Mokkink, Ellen B. M. Elsman, and Caroline B. Terwee. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Quality of Life Research*, 33(11):2929–2939, 2024. doi: 10.1007/s11136-024-03761-6.

[6] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

[7] Karel G. M. Moons, Johanna A. A. Damen, Tabea Kaul, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388:e082505, 2025. doi: 10.1136/bmj-2024-082505.

[8] Baptiste Vasey, Myura Nagendran, Bruce Campbell, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5):924–933, 2022. doi: 10.1038/s41591-022-01772-9.

[9] Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, 2025. doi: 10.1136/bmj-2024-081554.

[10] Trisha Greenhalgh, Joseph Wherton, Chrysanthi Papoutsi, et al. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, 19(11):e367, 2017. doi: 10.2196/jmir.8775.

[11] Laura J. Damschroder, Caitlin M. Reardon, Marilla A. Opra Widerquist, and Julie Lowery. The updated Consolidated Framework for Implementation Research based on user feedback. *Implementation Science*, 17:75, 2022. doi: 10.1186/s13012-022-01245-0.

[12] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. doi: 10.2307/2529876.

[13] Christopher H. Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–28, 2011. doi: 10.18637/jss.v038.i08.

[14] Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011. doi: 10.1111/j.1541-0420.2010.01546.x.

[15] Tal Yarkoni. The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, 44(2):180–198, 2010. doi: 10.1016/j.jrp.2010.01.002.

[16] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

[17] Kathryn Skivington, Lynsay Matthews, Sharon Anne Simpson, et al. A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ*, 374:n2061, 2021. doi: 10.1136/bmj.n2061.

[18] Bryce B. Reeve, Ron D. Hays, Jakob B. Bjorner, et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1):S22–S31, 2007. doi: 10.1097/01.mlr.0000250483.85507.04.

[19] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[20] I. O. Okpara, R. T. Abah, and M. O. Durojaiye. Numerical Computation and Epidemiological Risk Analysis of Sepsis Mortality among Adults with ICD-Coded Sepsis in MIMIC-IV. *International Journal of Innovative Science and Research Technology*, 11(8), 2026. doi: 10.38124/ijisrt/26aug1493.

[21] Jonathan A. C. Sterne, Ian R. White, John B. Carlin, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338:b2393, 2009. doi: 10.1136/bmj.b2393.

[22] Neil S. Jacobson and Paula Truax. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1):12–19, 1991. doi: 10.1037/0022-006X.59.1.12.

[23] David M. Eddy, William Hollingworth, J. Jaime Caro, Joel Tsevat, Kathryn M. McDonald, and John B. Wong. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices task force–7. *Value in Health*, 15(6):843–850, 2012. doi: 10.1016/j.jval.2012.04.012.

[24] I. O. Okpara, S. A. Kambai, and V. C. Obinna. A Fractional-Order Mathematical Model for Infectious-Disease Transmission and Control: Stability, Sensitivity and Numerical Analysis. *Asian Journal of Pure and Applied Mathematics*, 8(1):788–800, 2026. doi: 10.56557/ajpam/2026/v8i1303.

[25] Uwe Siebert, Oguzhan Alagoz, Ahmed M. Bayoumi, Beate Jahn, Douglas K. Owens, David J. Cohen, and Karen M. Kuntz. State-transition modeling: A report of the ISPOR-SMDM modeling good research practices task force–3. *Value in Health*, 15(6):812–820, 2012. doi: 10.1016/j.jval.2012.06.014.

[26] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1): 128–138, 2010. doi: 10.1097/EDE.0b013e3181c30fb2.

[27] Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.

[28] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11): e745–e750, 2021. doi: 10.1016/S2589-7500(21)00208-9.

[29] World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. World Health Organization, Geneva, 2021. ISBN 9789240029200.

[30] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, et al. The CARE principles for Indigenous data governance. *Data Science Journal*, 19:43, 2020. doi: 10.5334/dsj-2020-043.

[31] Enola Proctor, Hiie Silmere, Ramesh Raghavan, et al. Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(2):65–76, 2011. doi: 10.1007/s10488-010-0319-7.

[32] Joseph E. Alderman, Joanne Palmer, Elinor Laws, et al. Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, 2025. doi: 10.1016/S2589-7500(24)00224-3.

[33] Sally Hopewell, An-Wen Chan, Gary S. Collins, et al. CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, 389:e081123, 2025. doi: 10.1136/bmj-2024-081123.

[34] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, and SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26(9):1364–1374, 2020. doi: 10.1038/s41591-020-1034-x.