Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpreting Evaluation Indicators in Health Policy Implementation: A Critical Interpretive Synthesis Linking Indicators, Constraints and Strategies

Anny Mireilles Sarea de Mendonça

up202300377@up.pt

University of Porto https://orcid.org/0009-0005-9393-8712

Artur Rocha

Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência

José Coelho Rodrigues

University of Porto

Research Article

Keywords: Health policy implementation, Policy evaluation, Evaluation indicators, Efficiency, equity, and sustainability, Strategies and best practices, Health systems governance

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11693072/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Interpreting Evaluation Indicators in Health Policy Implementation: A Critical Interpretive Synthesis Linking Indicators, Constraints and Strategies

Abstract

Background When evaluation indicators show that a health policy is underperforming, those responsible must decide whether to adapt how it is executed, the resources behind it, or the policy instruments themselves. Evaluation evidence seldom indicates which. This study develops an interpretive framework linking evaluation indicators, reported constraints and reported responses to structure that judgement.

Methods We conducted a critical interpretive synthesis of 128 studies published between 2015 and 2025, purposively selected from Scopus, Web of Science and PubMed. Indicators were mapped to efficiency, equity and sustainability; constraints were classified as operational, institutional or structural; and reported responses were grouped into seven strategy categories. Patterns were compared across four income tiers and a multi-country or global category. Selection and coding were undertaken by the first author, without independent duplicate assessment or formal methodological quality appraisal.

Results The framework specifies a sequence: read the configuration across indicator dimensions, where some measures deteriorate while others hold; form competing hypotheses about which constraint type may be binding; investigate these against contextual information, which may rule them out; and only then consider responses. Two findings motivate it. Presence and breadth are not the same: sustainability was addressed in 106 of the 128 studies yet carried by the narrowest set of recurrent indicators, and equity by 59 studies with an intermediate set — a difference in how the included literature represents each dimension, not evidence that a

dimension is inadequately measured. Constraints rarely appeared alone: 60 of the 98 studies reporting a constraint described more than one type, so other types warrant investigation before responses are selected.

Conclusions The synthesis proposes that the configuration across dimensions is itself a signal about which kind of constraint to investigate, and organizes three questions before a policy is adapted: which explanations merit investigation, what information would distinguish them, and which responses are feasible under local conditions. Using evaluation data to guide action is established; the narrower claim concerns reading the configuration. The step from indicator pattern to constraint is a proposition developed by the authors, not a demonstrated capability; its accuracy, feasibility and effect on decisions require empirical evaluation.

Keywords: Health policy implementation; Policy evaluation; Evaluation indicators; Efficiency, equity, and sustainability; Strategies and best practices; Health systems governance

Background

The difficulty in policy implementation is often not a lack of indicators but that the indicators available are read as performance outcomes rather than as diagnostic signals: they record that something is not working, without explaining which implementation weaknesses are limiting performance, or which responses are worth considering. Across diverse institutional environments, the implementation of public health policies is reported to falter where the structural, institutional, and operational conditions needed for sustained execution are absent or misaligned [1, 2, 3]. This persistence of implementation failure, despite decades of health reform investment, points to a governance problem that this review takes as its starting point: the instruments used to detect failures, the weaknesses that may underlie them, and the strategies available to address them are studied separately, and that separation may limit how health systems learn and adapt.

Reviews of evaluation indicators, of implementation weaknesses and of mitigation strategies have largely developed as separate literatures, and the frameworks that span them are specified for other purposes, as the discussion sets out. This synthesis brings the three together within a single analytical framework derived from the accumulated evidence base, offered as a conceptual contribution rather than an empirically validated model. Low- and middle-income settings contribute the largest share of its evidence base, though not a majority of it: 60 of the 128 included studies (47%) describe implementation in low-, lower-middle- or upper-middle-income countries, against 40 (31%) in high-income or OECD settings, with the remaining 28 (22%) multi-country or global in scope and not assigned to a single income tier. The questions it addresses, how to read routine indicator data when supervision, financing and workforce constraints are reported together, and how to choose among responses under those conditions, are the questions those systems face most acutely; high-income evidence enters here as comparison rather than as the reference case against which other settings are read. The remainder of this Introduction situates that contribution against the existing literature.

The literatures on evaluation indicators, implementation weaknesses, and mitigation strategies in health policy are each well developed, which is precisely what makes a synthesis across them possible: the difficulty is not an absence of evidence but its separation. Studies of evaluation indicators have documented a persistent preference for efficiency-oriented measures, service utilization, coverage rates, and quality of care, at the expense of equity and sustainability [4, 5, 6]. Studies of implementation barriers have catalogued structural weaknesses such as underfunding and infrastructure fragility [7, 8, 9], institutional fragilities including governance fragmentation and weak accountability [10, 11], and operational failures such as weak supervision and implementation fidelity [12, 13]. Studies of mitigation strategies have evaluated performance-based financing [14, 15], governance reforms [16], community engagement [17,

18], and health information systems [19, 20]. What this body of work has given less attention to is how indicators already reported might be read together to frame questions about which constraints are operating and which responses to consider.

Three governance gaps follow directly. First, we are not aware of a comparative synthesis of which indicators are most recurrently used across different health system contexts, or of whether they capture efficiency, equity, and sustainability together [21, 6]. Second, implementation weaknesses are rarely analysed in relation to the indicators most sensitive to their presence, making it difficult for health managers to infer from performance data which weakness may underlie underperformance [21, 10]. Third, mitigation strategies are seldom linked systematically to the weakness categories they address or the indicators most appropriate to assess their effects [22, 23]. Addressing these three gaps together could inform governance decisions that fragmented reviews leave open: which indicators to select, which strategies to deploy, and how to evaluate their effects.

Existing frameworks address complementary components of this problem. CFIR and EPIS emphasize implementation determinants and processes [24, 25], while WHO health-system frameworks organize system functions and indicators [26, 27]. More integrative approaches, including CFIR-ERIC and the Implementation Research Logic Model, link determinants to strategies and outcomes [23, 28]. Building on these traditions, this synthesis sets out how indicator patterns, weakness profiles and mitigation strategies might be read together within a context-sensitive cycle.

This synthesis does not seek to identify which strategy performs best, but which strategies the literature reports in relation to a given weakness profile. It is guided by three research questions.

- **R.Q.1:** What are the most relevant and recurrent indicators used to measure health policy implementation success, and to what extent do they support evaluation of efficiency, equity, and sustainability?
- **R.Q.2:** What structural, institutional, and operational weaknesses are most frequently identified, and how do they vary across health system contexts?
- **R.Q.3:** What mitigation strategies are recurrently reported in relation to identified implementation weaknesses, and which indicators are used to evaluate their implementation and outcomes?

To address these questions, this study synthesizes 128 studies published between 2015 and 2025 through a critical interpretive synthesis, classifying the indicators, weaknesses, and strategies reported in each and comparing them across country income contexts. The findings show substantial contextual variation in indicator priorities, with equity the dimension least often addressed in every setting; frequent co-occurrence of implementation weaknesses; and recurrent associations between weakness profiles and the mitigation strategy categories reported alongside them. The comparison includes 60 studies from low- and middle-income settings (46.9% of the corpus), 40 from high-income settings and 28 multi-country or global studies; the latter are not assigned to a single income tier. Together these patterns provide the empirical basis for the integrated framework developed here, whose central contribution is to propose reading indicators prospectively, as pointers to what may need investigating, rather than only retrospectively, as a record of what has already happened.

Methods

This study is a critical interpretive synthesis (CIS) [29], an approach developed for reviewing large and heterogeneous literatures in which the review question and the analytical categories are allowed to develop through interpretive engagement with the evidence rather than being fixed in advance. It is not a systematic review and does not claim adherence to PRISMA. Purposive rather than exhaustive sampling, the absence of a pre-specified protocol, and the interpretive construction of categories are choices made in this study, consistent with the approach but not entailed by it: critical interpretive syntheses have been conducted with structured search procedures and with independent duplicate coding at each stage [30]. These choices are justified below, and their consequences are set out in the limitations. Justifying them does not show that they left the results unaffected; single-coder classification and the absence of critical appraisal remain sources of uncertainty that this design cannot quantify. What CIS does retain from systematic-review practice is transparency about how the evidence was found, selected and handled, and the search, selection and coding steps are reported here in those terms. A PRISMA 2020 checklist, completed to the extent applicable to an interpretive design, accompanies the submission as an aid to transparency rather than as a claim of adherence, indicating which items are reported and why the remainder do not apply to this design [31]. Searches were conducted in Scopus, Web of Science, and PubMed using terms combining health policy implementation, evaluation indicators, implementation weaknesses, and mitigation strategies (full search strings in Supplementary Appendix A), without date or language filters. The lower temporal bound of 2015 was established to capture evidence generated since the adoption of the Sustainable Development Goals framework [32, 33], which substantially reoriented global health policy evaluation towards the equity and sustainability dimensions central to this synthesis. Inclusion criteria required studies to be published between 2015 and 2025 in English or Portuguese, address health policy implementation or evaluation, report indicators relating to efficiency,

equity or sustainability, or discuss implementation weaknesses or mitigation strategies. Studies focused exclusively on clinical efficacy without implementation relevance were excluded.

Consistent with the interpretive orientation of CIS, no rigid a priori protocol or fixed review question was specified in advance; the analytical categories emerged from iterative engagement with the literature [29].

The database search yielded 941 records. After removal of duplicates (n = 316), 625 records remained; following language exclusions (n = 25), 600 titles and abstracts were screened, and 200 reports were purposively selected for full-text assessment. Selection at this stage applied four operational criteria: conceptual relevance to at least one of the three analytical dimensions; contribution to an analytical category that was still developing; diversity of health-system contexts, so that no income tier or region dominated the emerging categories; and retention of contradictory or deviant cases that challenged emerging patterns. The remaining 400 records were not taken forward. Consistent with the purposive logic of CIS, they were passed over because they did not add to a developing analytical category, rather than being excluded one at a time against a fixed checklist; no count by reason for exclusion is therefore available at this stage, which is a property of the sampling approach rather than an omission in its reporting. At full-text assessment, 72 were excluded (published before 2015: n = 58; no implementation-relevant evaluation dimensions: n = 12; outside the health policy domain or lacking implementation-relevant content: n = 2), yielding 128 included studies. Sampling within each analytical category continued until additional records yielded no new categories or properties, judged interpretively rather than by a pre-specified stopping rule. The resulting body of evidence is a purposively assembled sample designed to support interpretive synthesis, not a probability sample: the frequencies reported below describe this body of evidence rather than estimating prevalence in the wider literature. The 128 included studies are listed individually in the

reference list [1–18, 20–22, 26, 34–139] and are mapped to study characteristics, income group, and extracted categories in Supplementary Appendix I.

The database searches were run in August 2025. Title-and-abstract screening, purposive selection, full-text assessment, data extraction and coding were carried out by the first author. No independent duplicate screening or coding was undertaken. The eligible publication window runs from 2015 to the end of 2025, although the corpus is in practice bounded by the August search date. One included study [51] was published online in September 2025 and therefore did not arise from the database searches; it was retained because it met the inclusion criteria.

Indicators were classified across three evaluative dimensions: efficiency (service utilization, coverage, quality, protocol adherence and, where reported, cost and timeliness); equity (access disparities, financial protection, vulnerable population coverage); and sustainability (financing continuity, governance quality, workforce resilience, interoperability). The three labels follow the evaluation vocabulary used in the included studies. Efficiency is accordingly the broadest of the three and is not used here in its economic sense of a ratio between resources and outputs; what the category contains is reported in the Results and set out indicator by indicator in Supplementary Appendix C. An indicator was recorded whenever a study reported it, whether it was measured in that study, specified in a protocol, or proposed in a conceptual framework; the extraction did not distinguish these states, so the counts below describe what the literature reports rather than what has been measured. Indicators could be mapped to more than one dimension; where a single primary dimension was required, the dominant analytical purpose was used. Indicators were extracted from the full text of every study and verified study by study: all 128 studies report evaluation indicators. Extracted formulations were grouped thematically where they referred to the same underlying measure. Recurrence was assessed by the number of included publications reporting each indicator, after grouping formulations referring to the

same underlying measure. This frequency-based approach is consistent with previous reviews identifying commonly reported health indicators [140, 141]. In this synthesis, indicators reported in at least six of the 128 publications (4.7%) were classified as recurrent, yielding the 45 reported here; the full set, with operational interpretations and assigned dimensions, is given in Supplementary Appendix C. The threshold is a study-specific operational choice and not a validated standard: Gathara et al. [140], for example, treated indicators appearing in four or more studies as commonly reported. Recurrence describes repeated reporting within the included literature; it does not establish an indicator's importance, validity or suitability for a particular context. Because an indicator may address more than one dimension, the dimensional proportions reported below (efficiency 41%, equity 31%, sustainability 28%) are percentages of the 75 indicator-to-dimension assignments, not of the 45 indicators or the 128 studies. Implementation weaknesses were grouped into structural, institutional, and operational categories, defined in Supplementary Appendix D, and mitigation strategies into seven recurrent categories, each examined for the weakness categories alongside which it was reported. Where this article refers to constraints on implementation, it means these coded weakness categories; the two terms are used interchangeably. Weaknesses and strategies were coded study by study: a category was assigned only where the study itself reported it, and the supporting wording was recorded so that the coding can be checked and independently replicated. Studies reporting no weakness were retained in the synthesis but excluded from the weakness and strategy counts, which therefore rest on 98 studies rather than 128. The unit of analysis is the publication rather than the underlying programme: evidence reported in a primary study may also appear within an included review, so counts should be read as the number of publications reporting a given category.

Each study was assigned to one country income group using the World Bank classification for fiscal year 2024: low-income (LIC), lower-middle-income (LMIC), upper-middle-income (UMIC), or high-income/OECD. Studies whose countries all fell within a single tier were assigned to that

tier (LIC, n = 14; LMIC, n = 29; UMIC, n = 17; high-income/OECD, n = 40), while studies spanning more than one tier, or global in scope, were classified as multi-country/global (n = 28). This last category comprises analyses of several countries in different tiers (n = 11), regional studies (n = 12), and global analyses or reviews without a defined geographical base (n = 5); it is a residual category defined by the absence of a single income context. Criteria are given in Supplementary Appendix B and the study-level classification in Supplementary Table S5. Because income tiers change annually, one classification year was applied throughout; under FY2025, the five Iranian studies would move to UMIC.

Given the heterogeneity of study designs across the 128 included studies, spanning randomized evaluations, quasi-experimental analyses, observational studies, programme evaluations, and systematic reviews, no formal critical appraisal of methodological quality was undertaken. Although design-specific appraisal tools such as MMAT, AMSTAR 2, or GRADE could have been applied to subsets of the included studies, such assessments would not have provided a common evidentiary metric aligned with the purpose of this synthesis. Heterogeneity of design is, on its own, an incomplete justification, since appraisal tools exist for each design represented and could have been applied within design strata. The decision rests instead on what an appraisal score would have been used for here. No study is weighted, pooled or ranked in this synthesis, and studies enter it as accounts of how indicators, weaknesses and strategies were reported together rather than as estimates of effect; an appraisal score would therefore not have altered which studies informed which analytical category. Consistent with the interpretive orientation of CIS, studies were included and interpreted according to their relevance to the developing constructs rather than weighted by an appraisal score, and no study was excluded on quality grounds. The cost of that decision falls on the conclusions rather than on the analysis, and we state it directly: the associations reported between weakness profiles and strategy categories reflect what studies of varying robustness reported alongside one another, and cannot be read as evidence that those strategies were effective.

Use of generative AI tools

During manuscript preparation, the authors used Claude (Anthropic, Claude Opus 5; February–October 2026) for structural feedback, drafting and rephrasing selected passages, journal formatting, reference-style conversion, figure generation (Figures 1–4 and the study-selection diagram), dataset assembly from author-verified tables, and consistency checks. ChatGPT (OpenAI, GPT-4) and Grammarly Pro for Windows were used for language editing and English translation (August–October 2026).

All AI-assisted outputs were reviewed, edited and verified by the authors against the original studies, coded dataset and supplementary materials, as applicable. AI tools were not used for literature searches, study screening or selection, data extraction, coding, development of analytical categories or framework construction; these tasks were performed by the first author. The authors retain full responsibility for the manuscript and its AI-assisted content. No AI tool is listed as an author.

Results

The synthesis is directed at a particular interpretive task: starting from the patterns a system already reports in its evaluation indicators, and asking which implementation constraints might account for them and which responses the literature reports alongside those constraints. The findings below are organized around three interconnected elements that this task requires: the indicator groups most recurrently used across health system contexts; the implementation weakness profiles that most frequently constrain policy execution; and the mitigation strategies reported in relation to those weaknesses, together with the indicators used to evaluate them. These findings constitute the empirical basis for the integrated framework presented at the end of this section. The three analyses do not share a denominator: indicator reporting was verified

in all 128 studies, while 98 studies (77%) reported an identifiable implementation weakness and those same 98 reported at least one mitigation strategy.

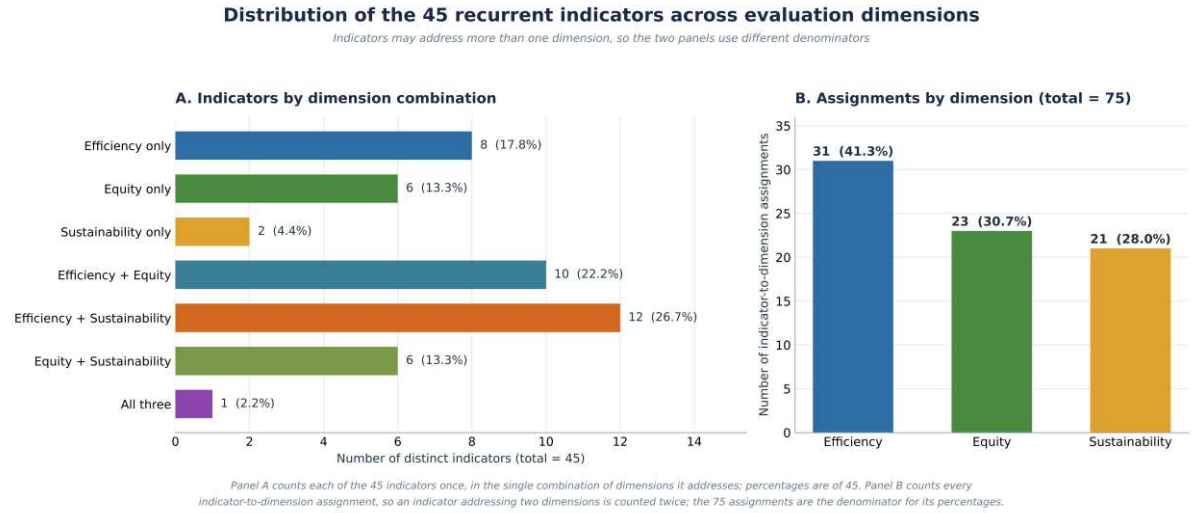
Indicators (R.Q.1)

Forty-five recurrent indicators were identified across the 128 included studies (see Figure 1). At the indicator level, efficiency was the most widely covered dimension. Across the 75 indicator-to-dimension assignments, efficiency accounted for 31 assignments (41%), equity for 23 (31%), and sustainability for 21 (28%). Because individual indicators could map to more than one dimension, these assignments corresponded to 31, 23, and 21 of the 45 recurrent indicators, respectively, and therefore sum to more than 45 (full indicator set in Supplementary Appendix C). What the efficiency category contains bears stating, because it governs how that 41% should be read. Of the 31 recurrent indicators assigned to efficiency, three are cost measures (cost-effectiveness and avoided expenditure, hospital costs, and cost per patient or service) and two concern timeliness or service load; the remaining 26 are measures of service delivery and performance, principally utilization, coverage, quality, protocol adherence and health outcomes. The predominance reported here is therefore a predominance of service-performance indicators rather than of efficiency indicators in the economic sense. Under the classification adopted, equity and sustainability were represented by smaller sets of recurrent indicators; because the efficiency category is the broadest of the three, part of that difference follows from how the categories were defined rather than from what the literature attends to. The study-level frequencies reported below are a separate analysis and should be read separately. At study level, the pattern differs for sustainability, which was addressed by 106 of the 128 studies (83%; e.g. [8, 60, 114, 3]) and efficiency by 96 (75%; e.g. [47, 108, 5, 139]), while equity was addressed by only 59 (46%; e.g. [76, 15, 125, 128]). The two levels therefore give different orderings, and the distinction matters for interpretation. A narrower set of recurrent indicators is not in itself evidence that a dimension is inadequately measured, since a small set

of well-chosen measures may be sufficient; what it shows is that fewer distinct measures recurred across this corpus. At indicator level, sustainability is carried by the narrowest set of recurrent indicators (21 of 45), with equity intermediate (23) and efficiency widest (31); at study level, equity is the dimension least often addressed, by 46% of studies against 75% for efficiency and 83% for sustainability. The two counts therefore describe two different gaps: sustainability is widely treated as a concern but translated into comparatively few distinct indicators, whereas equity is carried by a moderate number of indicators that fewer than half the studies report. Throughout, a dimension is counted as addressed by a study when that study reports at least one indicator mapped to it; this records what was reported, not how well it was measured. The most frequent cross-dimensional combination was Efficiency + Sustainability (26.7%), followed by Efficiency + Equity (22.2%).

Figure 1: Distribution of the 45 recurrent indicators across evaluation dimensions

Alt text: Two-panel bar chart. The left panel shows seven mutually exclusive combinations of evaluation dimensions across 45 recurrent indicators, with counts of 8, 6, 2, 10, 12, 6 and 1. The right panel shows 31 efficiency, 23 equity and 21 sustainability assignments; percentages use the 75 indicator–dimension assignments as denominator.



Equity was the dimension least often addressed at study level in all contexts, and sustainability, although frequently invoked as a concern, was carried by the narrowest set of standalone indicators. Indicator priorities varied by country income group (Supplementary Table S4 and Supplementary Figure S1). In LIC settings, the most frequently reported indicator themes were governance and accountability, service coverage and utilization, including maternal and child health services [142], and quality of care and protocol adherence. In LMICs, service coverage and utilization predominated, followed by financing continuity and sustainability, workforce availability and retention, and health outcomes. In UMICs, costs and efficiency were the most frequently reported theme, followed by service coverage and utilization and health outcomes. In high-income/OECD settings, service coverage and utilization, quality of care and protocol adherence, and user experience and satisfaction were reported most often.

The same indicators, however, were reported alongside different implementation weaknesses depending on context. Structural weaknesses, for instance, were reported proportionally more often the lower the income tier: in 8 of the 10 low-income studies reporting a weakness, 18 of 26 lower-middle-income and 7 of 11 upper-middle-income, against 15 of 28 in high-income settings (Supplementary Table S4). These are descriptive counts within a purposively assembled corpus, not tested differences, but they indicate that which weakness accompanies a given indicator is not fixed across settings, and that indicator performance is not a direct, context-independent readout of implementation quality. The following section therefore examines which weaknesses these indicator patterns most frequently accompanied in the reviewed studies, and whether those weaknesses co-occurred in ways that a single indicator dimension could not reveal on its own.

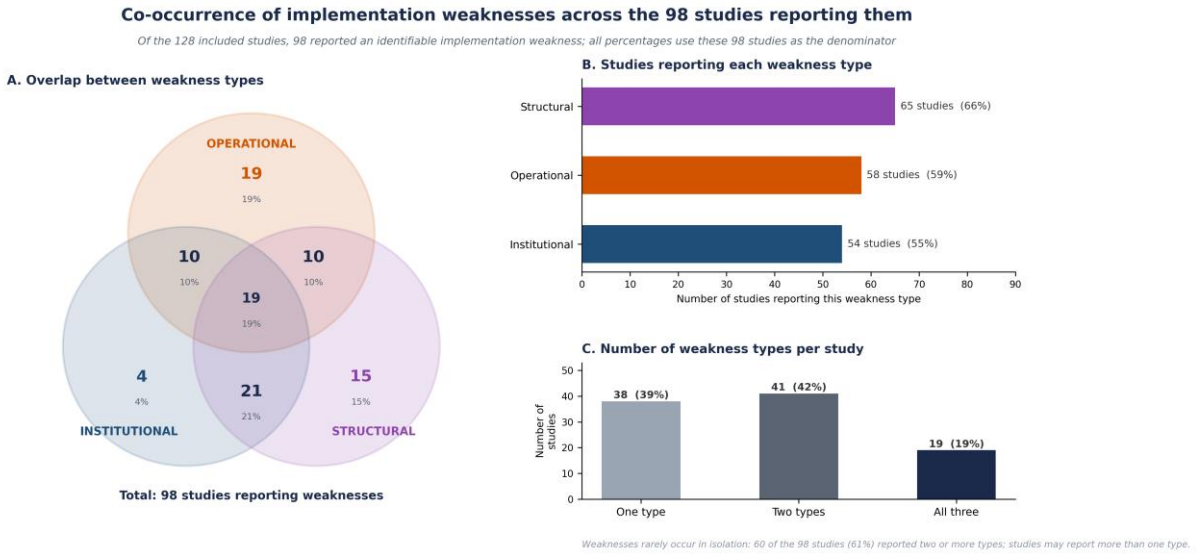
Implementation weaknesses (R.Q.2)

Of the 128 studies, 98 (77%) reported an identifiable implementation weakness; the remaining 30 were study protocols, conceptual papers, or evaluations reporting only positive outcomes

(e.g. [42, 69, 103, 116]), and are therefore excluded from the weakness counts. All proportions reported in this section use these 98 studies as the denominator. Structural weaknesses were the most frequently identified within the sampled body of evidence, appearing in 65 studies (66%; e.g. [50, 73, 15, 125]), followed by operational (58 studies, 59%; e.g. [4, 14, 98, 118]) and institutional (54 studies, 55%; e.g. [55, 86, 92, 111]) (see Figure 2). Operational weaknesses typically involved supervision gaps [143], limited technical capacity, and implementation fidelity problems at the frontline of service delivery ([1, 14, 13, 118]). Institutional weaknesses centred on governance and coordination failures, weak accountability, and poor intersectoral integration ([55, 10, 100, 11]), while structural weaknesses involved financing constraints, infrastructure deficits, and weak information systems ([7, 57, 122, 131]). Weaknesses often co-occurred: while 38 studies (39%) reported a single weakness type, 41 (42%) reported two and 19 (19%) reported all three (e.g. [20, 105, 112, 136]).

Figure 2: Co-occurrence of implementation weaknesses across the 98 studies reporting them

Alt text: Three-panel figure based on the 98 studies reporting an implementation weakness. Panel A is a Venn diagram of operational, institutional and structural weaknesses: 19 studies operational only, 4 institutional only, 15 structural only, 10 operational with institutional, 10 operational with structural, 21 institutional with structural, and 19 reporting all three. Panel B is a horizontal bar chart of studies reporting each type: structural 65 (66%), operational 58 (59%) and institutional 54 (55%). Panel C is a vertical bar chart of the number of weakness types per study: one type 38 studies (39%), two types 41 (42%) and all three types 19 (19%).



355

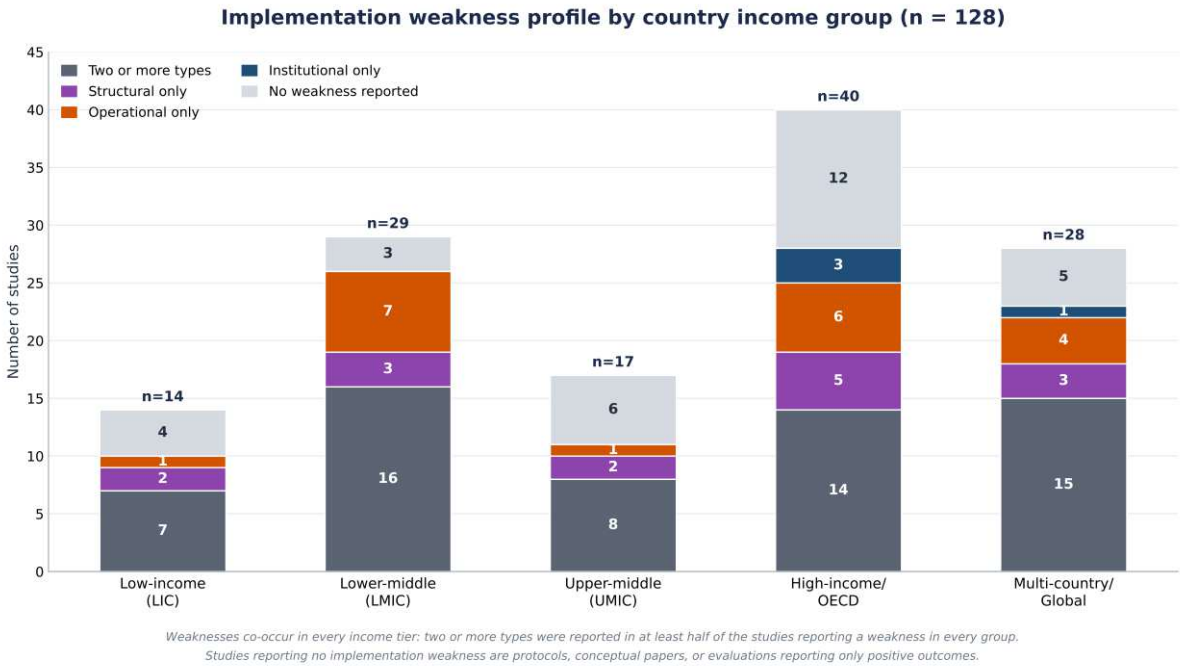
356 Co-occurrence was found in every income group rather than being a feature of any one tier: two
357 or more weakness types were reported in at least half of the studies reporting a weakness in
358 each group, namely 7 of 10 in LIC settings, 16 of 26 in LMICs, 8 of 11 in UMICs, 14 of 28 in high-
359 income/OECD settings, and 15 of 23 in multi-country or global studies. Structural weaknesses
360 were reported across all income groups, including high-income settings, where they took the
361 form of financing pressures and workforce constraints rather than absolute scarcity; the profile
362 therefore differentiates less by income tier than by the specific combination of weaknesses
363 reported in each setting. The most frequent single-type profiles were operational-only (19
364 studies, 19%) and structural-only (15, 15%), with institutional-only rarer (4, 4%), while the most
365 common pairing was institutional with structural (21, 21%); the full distribution is shown in
366 Figure 2 and Figure 3.

367 The dominant weakness profile of a health system therefore organizes, within the framework,
368 which indicators are treated as most relevant to examine and which strategy categories are
369 brought into consideration. This should not be read as implying that a single weakness type acts
370 alone: since 60 of the 98 studies (61%) recorded two or more types, dominance denotes the
371 weakness reported as most binding in a given setting, not the only one present, and where no

type is clearly dominant the compound profile applies, so multiple or sequenced strategies may warrant consideration rather than assuming a single strategy is sufficient.

Figure 3: Implementation weakness profile by country income group

Alt text: Stacked vertical bar chart of the 128 included studies by country income group. Each bar is divided into studies reporting two or more weakness types, structural only, operational only, institutional only, and no weakness reported. Low-income (n = 14): 7, 2, 1, 0 and 4. Lower-middle-income (n = 29): 16, 3, 7, 0 and 3. Upper-middle-income (n = 17): 8, 2, 1, 0 and 6. High-income/OECD (n = 40): 14, 5, 6, 3 and 12. Multi-country or global (n = 28): 15, 3, 4, 1 and 5. Two or more weakness types were reported in at least half of the studies reporting a weakness in every group.



Mitigation strategies (R.Q.3)

Seven mitigation strategy categories were identified across the 98 studies reporting an implementation weakness (definitions and evidence in Supplementary Appendix E) (see

Supplementary Figure S2). All 98 reported at least one category, and most reported more than one: 21 studies reported a single category, 54 reported two, and 23 reported three or more, giving 198 strategy-category assignments in total. The 30 studies without an identified weakness reported interventions, protocols, or conceptual models, but no mitigation of a reported weakness. Counts below are numbers of studies and are not mutually exclusive.

"Capacity Building and Qualification" (44 studies, 45%) was most often reported against operational weaknesses; "Governance and Leadership Strengthening" (42, 43%) against institutional weaknesses; "Funding Reforms and Incentives" (31, 32%) against structural weaknesses; "Information Systems and Indicators" (27, 28%) disproportionately against operational weaknesses; "Organizational Reforms and Service Integration" (27, 28%) where services were fragmented; "Community Engagement and Co-Design" (21, 21%) against trust, adherence and equity-related weaknesses; and "Infrastructure, Technology and Innovation" (6, 6%) against geographic and structural weaknesses. Full definitions, evidence and the indicators used to assess each category are given in Supplementary Appendix E.

Weaknesses, strategies, and indicators also did not appear as independent elements across the reviewed studies: particular weakness profiles were repeatedly discussed together with particular mitigation strategies and evaluated through specific indicators. Operational weaknesses, for example, were frequently addressed through capacity-building interventions and subsequently monitored using indicators such as protocol adherence, workforce readiness, and service utilization.

Table 1 summarizes how the weakness categories, the mitigation strategy categories reported alongside them, and the evaluation indicators used relate to one another; the expanded matrix is given in Supplementary Appendix F.

Table 1: Weakness categories with the mitigation strategies and evaluation indicators most frequently reported alongside them.

Weakness category (studies)	Most frequently reported strategies	Indicators used to evaluate them	Dimension emphasized
Operational weaknesses (58 studies)	"Capacity Building and Qualification" (30 of 58); "Governance and Leadership Strengthening" (23); "Information Systems and Indicators" (22) [14, 20, 70, 112]	Protocol adherence; quality of care; service utilization; continuity of care	Efficiency
Institutional weaknesses (54 studies)	"Governance and Leadership Strengthening" (34 of 54); "Capacity Building and Qualification" (24); "Funding Reforms and Incentives" (20) [61, 70, 105, 136]	Interoperability and HIS; institutional capacity; programme continuity; referral efficiency	Sustainability
Structural weaknesses (65 studies)	"Governance and Leadership Strengthening" (31 of 65); "Capacity Building and Qualification" (28); "Funding Reforms and Incentives" (28) [41, 70, 105, 136]	Accessibility; urban-rural inequality; financial protection; workforce readiness	Equity / sustainability
Multidimensional weaknesses (60 studies with two or more types)	"Governance and Leadership Strengthening" (33 of 60); "Capacity Building and Qualification" (27); "Funding Reforms and Incentives" (22) [61, 70, 105, 136]	Indicators spanning two or more dimensions	Efficiency / equity / sustainability

The coded unit throughout is the publication, and a weakness category recorded for a publication is not equivalent to the dominant profile of the health system it describes: a study may report an operational weakness without having examined institutional or structural conditions. The framework proposes that such a profile be investigated in a setting; the synthesis does not establish it. For framework application, a dominant weakness profile denotes the weakness interpreted as most salient in a given setting; study-level coding itself was non-mutually exclusive, and cross-cutting denotes studies where operational, institutional, and structural weaknesses co-occur. A single weakness category was assigned only where the study itself characterized one weakness as primary, central, or binding; where no such characterization was made, and more than one category was reported, the study was classified as compound.

Cross-context patterns: regularities across country income groups

Four regularities hold across the four country income tiers and the multi-country or global category, and they motivate a single integrated framework rather than one per context. First, the lower frequency with which equity was addressed relative to efficiency, and the gap between how often sustainability was invoked and how few distinct indicators carried it, were observed in every group: none evaluates the three dimensions in balance, even though the indicator themes reported most often differ between groups (Supplementary Table S4). Second, weaknesses co-occur rather than acting in isolation in every context: structural weaknesses recurred across tiers and, where they co-occurred with another type, were more often reported alongside institutional than operational weaknesses in every group, and in no context did weaknesses reduce to a single, cleanly separable category. Third, mitigation strategy categories were themselves rarely reported in isolation: 77 of the 98 studies reporting a weakness reported two or more, mirroring the co-occurrence observed among the weaknesses. “Organizational Reforms and Service Integration” was the category most evenly reported across the three weakness categories (17 operational, 17 institutional, 16 structural), followed by “Capacity Building and Qualification”, while “Information Systems and Indicators” was reported disproportionately alongside operational weaknesses (22 of 27). Fourth, the synthesis proposes a common analytical structure linking indicators, implementation weaknesses and reported strategies across contexts. Contextual variation motivates adaptation of its contents; whether the proposed diagnostic logic is transferable across settings remains an empirical question. The framework is therefore a proposition for investigation rather than a demonstrated cross-context instrument.

Supplementary Table S4 consolidates indicators, weaknesses, and mitigation strategy categories across the five contexts, with the underlying counts by income group in Supplementary

Appendix G. The multi-country and global studies (28 studies; 23 reporting a weakness) are not anchored in a single income context: governance strengthening is their most frequently reported response (11 of 23), structural weaknesses were reported in 17 of 23, and efficiency is addressed by 68% of them against 79% for sustainability. Their profile is best read as reflecting an analytical vantage point across systems rather than a single health-system context.

This carries a practical implication for how evaluation frameworks travel between settings. Service coverage and utilization indicators were reported in every group, while the other most frequent themes differed: governance and accountability in LIC settings, financing continuity and workforce in LMICs, costs and efficiency in UMICs, and quality of care and user experience in high-income settings. The variation is one of emphasis rather than kind, but a framework transposed unchanged may still be less sensitive to weaknesses more frequently reported elsewhere, such as supervision, workforce, financing, and infrastructure constraints in lower-income settings. The reviewed evidence suggests that what may need adapting between contexts is the indicator set used to read the system; whether the diagnostic logic itself travels unchanged is a question this synthesis raises rather than one it settles. A cross-tabulation of the coded categories across the 98 publications reporting a weakness shows considerable overlap rather than exclusive correspondence: all three weakness types recur among the publications reporting each evaluation dimension, and all three dimensions recur among those reporting each weakness type, with no pairing confined to a single category (study-level classifications in Supplementary Appendix I). These publication-level co-occurrences do not establish whether the categories are statistically associated and cannot assess the propositions in Table 2, which concern joint indicator trajectories within a setting. Such trajectories were not captured by the coding used in this synthesis. Table 2 presents author-developed propositions for contextual investigation: strategy frequencies derive from the included studies, whereas indicator patterns, hypothesized weakness profiles and proposed monitoring indicators are interpretive constructions that have not been empirically tested.

475 Table 2: Analytical propositions for contextual investigation.

Indicator pattern[a]	Weakness profile[a]	Strategies reported[b]	Indicators for monitoring[a]	Alternative explanations and contextual checks[a]
Service-performance measures deteriorate while sustainability and equity measures remain stable (e.g. falling protocol adherence and service utilization, rising waiting time)	Operational: supervision gaps, protocol non-adherence, workforce turnover, weak monitoring routines [4, 14, 98, 118]	“Capacity Building and Qualification” (30 of 58); “Governance and Leadership Strengthening” (23); “Information Systems and Indicators” (22)	Protocol adherence; quality of care; continuity of care	Case-mix or demand change, or a change in recording practice, rather than a delivery failure. Check volumes and case-mix by period, whether the fall appears across facilities or only where recording changed, and the stability of denominators.
Sustainability and coordination measures deteriorate while service-performance measures hold (e.g. falling interoperability, referral efficiency and programme continuity)	Institutional: governance fragmentation, weak accountability, poor intersectoral coordination [55, 86, 92, 111]	“Governance and Leadership Strengthening” (34 of 54); “Capacity Building and Qualification” (24); “Funding Reforms and Incentives” (20)	Interoperability; governance quality; referral efficiency; programme continuity	A lag rather than a different profile: service performance may be held up by staff effort or buffer stocks and fall later. A system or platform migration can also depress interoperability measures without governance change. Check vacancy, overtime and stock buffers, and whether the decline coincides with a known system change.

Equity and access measures deteriorate, often with financing and coverage measures (e.g. falling territorial coverage, financial protection and workforce availability)	Structural: chronic underfunding, workforce shortages, infrastructure fragility, supply-chain instability [15, 50, 73, 125]	“Governance and Leadership Strengthening” (31 of 65); “Capacity Building and Qualification” (28); “Funding Reforms and Incentives” (28)	Accessibility; territorial coverage; financial protection; system resilience	Population change shifting the denominator, or improved identification of underserved groups making existing disparities newly visible. Check population estimates, whether disaggregation methods changed, and absolute alongside relative measures.
Several dimensions deteriorate together, with none clearly dominant	Compound or cross-cutting: two or more weakness types co-occurring (60 of the 98 studies reporting a weakness: 41 reported two types and 19 all three) [6, 20, 105, 112]	Combinations of strategy categories, most often “Governance and Leadership Strengthening” (33 of 60), “Capacity Building and Qualification” (27), and “Funding Reforms and Incentives” (22)	Multidimensional indicators spanning efficiency, equity, and sustainability	A single upstream shock — a funding interruption, an epidemic, industrial action — or a disruption of the data system, producing simultaneous movement without a compound profile. Check whether onset is datable and simultaneous, whether a documented event coincides, and whether measures recover together.

Robustness of the reported patterns

Because the body of evidence includes both primary studies and evidence syntheses, and evidence reported in a primary study may also appear within an included review, all principal results were recalculated with the eight reviews excluded (n = 120). The proportion of studies reporting an implementation weakness was unchanged at 77%; the relative ordering of weakness types was preserved (structural 67%, operational 58%, institutional 54%, against 66%, 59% and 55% across all 128 studies); co-occurrence was 61% in both; and the dimensional profile was stable (sustainability 84% against 83%, efficiency 74% against 75%, equity 46% in both). Capacity building, governance strengthening, and funding reforms remained the three most frequently reported strategy categories, with capacity building and governance strengthening tied in first place (40 studies each), while information systems moved ahead of organizational reform in fourth position (27 against 23 studies). The patterns described in this section therefore held when those eight reviews were removed. This does not exclude other forms of overlap between studies, which the design cannot rule out.

The integrated analytical framework

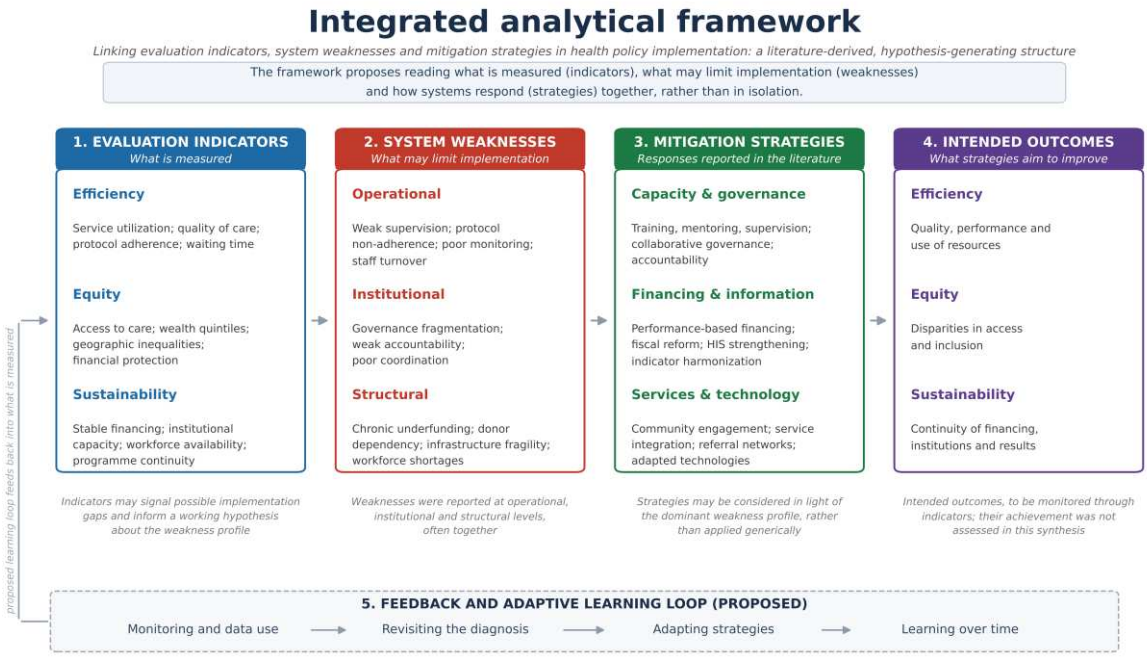
The relationships set out in this section describe how the framework is intended to be applied and the patterns observed across the reviewed studies; they are analytical propositions, not empirically established or statistically tested associations. In the reviewed studies, indicators, weaknesses, and strategies were frequently reported together rather than independently, which is consistent with viewing implementation performance as an interaction between what is reported, what constrains implementation, and how systems respond. Established frameworks organize parts of this space for other purposes, as the discussion sets out [24, 25, 26, 27].

Based on the patterns identified across the synthesis, we developed a framework that links three key elements: (1) indicators, (2) implementation weaknesses, and (3) mitigation strategies, arranged as a logical sequence rather than three independent lists (see Figure 4). First, indicators may provide signals of potential implementation problems. Second, these signals can inform a working hypothesis about the dominant weakness profile of a health system: operational, institutional, structural, or compound. Third, that hypothesis frames which strategy categories might be examined, drawing on those reported alongside that weakness profile rather than on a generic best-practice list; what the literature reports alongside a profile is a starting point for contextual appraisal, not an indication that a response suits the setting or has worked. Fourth, a strategy that is adopted is monitored through the indicators most suited to that strategy, closing the loop back to measurement. Fifth, the resulting, updated indicator signals feed into the next diagnostic cycle, so that the diagnosis can be revisited rather than reassessed only at fixed reporting intervals. In this sense, the framework proposes treating evaluation as part of a continuous learning cycle rather than as a retrospective reporting exercise.

Figure 4: Integrated analytical framework linking indicators, weaknesses, and mitigation strategies

Alt text: Diagram of the integrated analytical framework, read left to right across four linked columns. Column 1, evaluation indicators (what is reported), lists efficiency, equity and sustainability indicators. Column 2, system weaknesses (what may limit implementation), lists operational, institutional and structural weaknesses. Column 3, mitigation strategies (responses reported in the literature), groups them as capacity and governance, financing and information, and services and technology. Column 4, intended outcomes (what strategies aim to improve), restates efficiency, equity and sustainability. Below the four columns, a proposed feedback and adaptive learning loop runs from

monitoring and data use through revisiting the diagnosis and adapting strategies to learning over time, and feeds back into column 1. The framework is literature-derived and hypothesis-generating; the intended outcomes were not tested in this synthesis.



It is important to emphasize that this framework is not presented as a validated implementation model. Rather, it is a synthesis derived from recurring patterns identified across the reviewed literature. Its contribution is not the creation of new indicators or new implementation strategies. Instead, it provides an integrated analytical structure capable of linking what systems measure, what constrains implementation, and how policy responses are selected and evaluated. The framework should therefore be understood as a conceptual tool derived from the literature and intended to support future empirical work, rather than as an empirically validated causal model; its value lies in structuring governance decisions, not in predicting outcomes.

In the framework, these indicator patterns are treated as entry points for investigation: underperformance in efficiency indicators with stable sustainability measures may point to

operational weaknesses; deterioration in sustainability indicators with stable operational performance may point to institutional weaknesses; and simultaneous underperformance across multiple dimensions may point to structural or compound profiles. In each case the inference runs from the pattern across indicators, not from any single measure in isolation, and should be investigated against contextual knowledge of the system, and may be rejected by it, before a strategy is selected; three worked illustrations are given in Supplementary Appendix H. Updated indicator readings would then feed into the next cycle, which is intended as a continuous governance learning loop consistent with Learning Health Systems principles [68].

"Organizational Reforms and Service Integration" was, on qualitative synthesis of the included studies, the strategy category most evenly reported across all three weakness categories (17 operational, 17 institutional, 16 structural), followed by "Capacity Building and Qualification", indicating that these two were reported across operational, institutional, and structural profiles rather than tied to any one. This records the breadth with which they were reported, not demonstrated general applicability. "Information Systems and Indicators", by contrast, was reported disproportionately in relation to operational weaknesses (22 of the 27 studies reporting it also reported an operational weakness) and so did not, in these studies, function as the cross-cutting response its indicator focus might suggest. One qualification bears on this reading: because the synthesis is itself organized around evaluation indicators, the visibility of information-systems responses may partly reflect the orientation of the body of evidence rather than their intrinsic reach, so the distribution of strategies across weakness categories is advanced as a pattern warranting testing rather than as an established hierarchy. "Information Systems and Indicators" is nonetheless best understood as an enabling condition rather than a substitute for the resources it helps to allocate: the reviewed studies did not report HIS strengthening as a substitute for addressing chronic underfunding or workforce shortages; its reported role was to make those weaknesses visible [19, 20].

The dominance of efficiency-oriented indicators across all country contexts is consistent with Sun et al. [5] and Chai et al. [47]. This synthesis extends that evidence by documenting its cross-contextual consistency and adding a framework-level interpretation: the infrequency with which equity was addressed [32, 142], and the narrow indicator base carrying sustainability [3], may reflect not only a measurement gap but a possible misalignment between what health systems monitor and the dimensions most relevant to sustainable and equitable outcomes. The mapping of indicator priorities across income groups raises questions about applying standardized evaluation frameworks without contextual adaptation [87, 144].

Discussion

This section situates the findings in relation to existing implementation frameworks, considers what they imply for policy and management practice, and sets out the limitations that bound the interpretation offered here. The discussion returns throughout to the central observation of the synthesis: that indicators, weaknesses, and mitigation strategy categories are reported as interconnected rather than independent, and that this interconnection is what the integrated framework attempts to make usable.

Contributions to research

Several established frameworks already connect parts of this problem, and the contribution of the present synthesis is best stated by comparison with them rather than by claiming that the ground is empty. They differ from one another, and from the framework proposed here, along five axes: the unit of analysis; the relations they specify among determinants, strategies and outcomes; the role they assign to indicators; the attention they give to contextual adaptation; and what they leave to the user. CFIR [25] offers a determinant taxonomy for an intervention within an organizational setting, characterizing what may impede implementation without itself prescribing responses or measures. EPIS [24] organizes the same territory temporally, across

590 exploration, preparation, implementation and sustainment phases in public-service systems,
591 with sustained attention to outer-context influence. The WHO Building Blocks handbook [27]
592 does link system functions to indicators, and supplies measurement strategies for them; what it
593 does not do is read movement in those indicators back as evidence about which system function
594 is binding. UHC monitoring frameworks [26] likewise provide standardized indicators, with
595 progress rather than implementation constraint as the object. The two closest comparators are
596 the CFIR-ERIC matching tool [23], which maps identified determinants onto implementation
597 strategies, and the Implementation Research Logic Model [28], which specifies determinants,
598 strategies, mechanisms and outcomes together within a single model.

599 A second group of frameworks works at the level of policy and system rather than of a single
600 intervention, and these are the closer comparators for a synthesis of this kind. Bullock et al.
601 [145], also a critical interpretive synthesis, assembles eight determinant categories, a process
602 model and an inventory of strategies organized by governance, financial, service delivery and
603 implementation-support arrangements, each with an intended target. Kuchenmüller et al. [30]
604 identify six domains and five process stages for institutionalizing evidence-informed policy-
605 making, and supply staged descriptors against which an organization can judge how far
606 institutionalization has progressed. Varallyay et al. [146] set out a framework for evaluating
607 embedded implementation research, specifying stages, components and outcome types for the
608 evaluator rather than for the manager. Pacifico Silva et al. [147] propose nine dimensions across
609 five value domains against which a health innovation may be judged responsible. Each supports
610 a different task: cataloguing what may shape implementation at system scale; assessing the
611 maturity of a practice; designing an evaluation; and appraising an innovation normatively.

612 Against these, the framework proposed here differs in the direction of inference and in the data
613 it assumes, rather than in ambition. Its inference runs in the opposite direction from the
614 matching tools and from the determinant frameworks: rather than beginning from determinants

615 already identified and selecting strategies to match them, it begins from the indicator patterns
616 a system already reports and treats those patterns as grounds for a question about which
617 weakness profile may be present. None of these four frameworks, as published, sets out this
618 reading, although none excludes it: Varallyay et al. [146] note that their framework does not
619 articulate relationships between constructs or mechanisms of effect, and Bullock et al. [145]
620 state that determinants are hypothesized to affect outcomes without setting out the reverse
621 reading. Indicators are therefore an input to interpretation rather than only an outcome
622 measure, which is the role they carry in the monitoring frameworks and in the outcomes layer
623 of the policy-level frameworks. The second difference is the data the framework assumes: it is
624 specified for indicators a programme or system already reports. The comparators are specified
625 against information of other kinds — staged descriptors of institutionalization in Kuchenmüller
626 et al. [30], the constructs of an evaluation designed for the purpose in Varallyay et al. [146], the
627 value dimensions of an appraisal instrument under development in Pacifico Silva et al. [147],
628 and the determinant categories of an implementation analysis in Bullock et al. [145]. Whether
629 applying any of them requires information a system does not already hold will depend on the
630 setting. Contextual adaptation is approached comparatively, across four country income tiers
631 and a multi-country or global category, rather than through adaptation criteria applied case by
632 case.

633 Three of the included studies come closer still, and the comparison is sharper because they sit
634 inside the corpus this synthesis draws on. Durbin et al. [59] set out an implementation and
635 evaluation framework for a public dental programme, using a paediatric programme in Ontario
636 to show how specified goals and a monitoring strategy allow a programme to be adapted as data
637 accumulate. Requejo et al. [109], drawing lessons from a set of indicator-validation studies,
638 argue that accountability for maternal and newborn health requires monitoring bundles of
639 indicators that reveal how far policies are implemented and where law and practice diverge.
640 EQUIST, as applied by Uneke et al. [125] across six West African countries, moves from survey

data to action by locating the most deprived populations, identifying the bottlenecks they face, and estimating which intervention packages would avert the most deaths at what cost. The general idea that evaluation data should structure interpretation and guide action therefore has clear precedent, and this synthesis does not claim it. What distinguishes the reading proposed here is narrower and should be stated as such: none of the three treats the configuration across indicator dimensions, which measures fall while others hold, as the signal that indicates which kind of implementation constraint to investigate. Durbin et al. [59] specify indicators to track for one programme in one jurisdiction; Requejo et al. [109] specify the content of a measurement agenda for one health area; EQUIST reasons from equity gaps to intervention packages through modelled mortality and cost, which identifies who is underserved and what would help rather than why implementation is faltering. The proposition examined here is that the pattern itself carries diagnostic information, and it remains a proposition: the comparison establishes that it is distinct, not that it is correct.

Three differences bound what this synthesis can claim and are worth stating plainly. The first is provenance. Varallyay et al. [146] triangulated a retrospective inquiry, twenty-eight prospective key-informant interviews, fourteen expert interviews and a research-team survey; Kuchenmüller et al. [30] cross-checked a synthesis of eighty-eight documents against stakeholder dialogues with thirty-six participants from twenty country teams. The present synthesis draws on published literature alone, and the proposition that indicator patterns correspond to weakness profiles is a claim about regularities that literature synthesis cannot establish; it is for that reason set out as a question for contextual investigation rather than as a finding. The second is resolution. Three weakness classes are coarse beside eight determinant categories [145] or six domains [30], and where the institutional class absorbs governance, standards, leadership and organizational culture, the question it generates may be too broad to direct a response without further specification. The third is the strategy layer. The inventory in Bullock et al. [145] names instruments and their intended targets, where the categories used

here aggregate them; the two differ in granularity rather than in kind, and the choice between them depends on whether a user needs named instruments or a smaller set of categories. Equity is treated here as one of three indicator dimensions, which is a thinner treatment than applying it across the analysis, as Varallyay et al. [146] do through disaggregation and Pacifico Silva et al. [147] through a standing value domain.

The framework advances the literature in three ways. First, it integrates dimensions examined separately in prior reviews, indicators [4, 6], barriers [20, 3], and strategies [98, 23], within a single analytical architecture. Second, it brings transferability considerations [148, 149, 144] to bear at the level of practical governance decisions; the cross-context comparison raises questions of adaptation rather than demonstrating transferability. Third, it aligns with Learning Health Systems principles [68] by embedding evaluation within a continuous implementation-adaptation cycle. These converge in the operational decision cycle shown in Supplementary Figure S3: its diagnose-act-monitor loop specifies the data-to-decision logic of Learning Health Systems for the indicator-weakness-strategy relationship rather than for organizational learning in general, and its explicit revisit-the-diagnosis step renders adaptation a concrete decision rather than a general recommendation. It is the profile-matching step that distinguishes the cycle from a generic quality-improvement loop; the comparators discussed above specify that step differently, or locate it at the level of a single implementation effort rather than of the system.

Implications for policy and management

A key contribution of this synthesis is the repositioning of evaluation itself. Indicators are traditionally used for performance reporting, a retrospective account of whether targets were met, consistent with the efficiency-dominant use documented across national and cross-country health system evaluations [47, 108, 5]. The framework proposes a broader role: indicators could function as governance signals that support diagnosis and inform strategy selection and

adaptation. This shifts evaluation from retrospective accountability towards adaptive governance, applying the continuous-learning logic of Learning Health Systems [68] specifically to the indicator-weakness-strategy relationship.

The framework has implications for several stakeholders. For policymakers, it may support more context-sensitive policy design, since the dominant weakness profile of a health system, rather than a generic best-practice list, can help identify which strategic priorities and governance investments are worth considering, addressing a gap in the transferability literature where adaptation criteria exist but are rarely linked to a diagnostic profile [149, 144]. For managers, it offers a way to link observed indicator performance to a plausible weakness profile rather than treating underperformance as an undifferentiated signal. For researchers, it offers a shared vocabulary spanning indicators, weaknesses, and strategies. Structural weaknesses were more often reported with institutional than operational ones in every income group and were most often met by governance strengthening and funding reforms (Table 1), so these responses may warrant joint consideration.

For low- and middle-income health systems in particular, three features of the evidence bear directly on practice. First, compound weakness profiles were the norm rather than the exception in these settings: two or more weakness types were reported in 7 of the 10 low-income and 16 of the 26 lower-middle-income studies that reported any weakness, which argues against expecting a single strategy to suffice. Second, the indicator themes reported most often differ by tier, governance and accountability in low-income settings and financing continuity and workforce in lower-middle-income settings, so an evaluation framework imported unchanged from a high-income system may be least sensitive to the constraints these settings actually report. Third, because equity was the dimension least often addressed everywhere, and because the route to it proposed here is disaggregation of measures already collected rather than a new indicator set, it is an adjustment available even where measurement capacity is constrained.

Each of these three turns on reading data already reported in a different way. Whether they require new data collection depends on whether the relevant variables are already recorded, at sufficient quality, and at the level of disaggregation needed.

Operationally, the framework maps onto three stages of the policy cycle. At formulation, indicator patterns would anchor an initial working hypothesis; country-context patterns offer only background, since weakness profiles were nearly even across types in most income groups. At implementation, the proposed profile frames which of the seven strategy categories may be considered, and in what combination, so that choices rest on the weaknesses the data suggest rather than on those a generic toolkit assumes. At adaptation, the strategy-specific indicators are monitored; when they fail to move as expected, this may indicate that the diagnosis needs revisiting rather than that the strategy should be pursued for longer.

In practice, a manager would apply the framework in five steps: (1) read the current indicator signals for the service or system and check the quality of the underlying data; (2) with Supplementary Table S4 as background, form competing working hypotheses about the dominant weakness profile (operational, institutional, structural, or compound), together with the alternative explanations set out in Table 2; (3) investigate those hypotheses against contextual information, which may rule some of them out; (4) consider the strategy categories most often reported in Table 1 for whichever profile or profiles that investigation leaves standing — which may be more than one, and may remain uncertain — rather than a generic toolkit, and weigh them against local conditions; and (5) monitor the strategy-specific indicators and, if they do not move as expected, return to step 2 and revise the interpretation rather than persisting with the same strategy. Indicators open the investigation; they do not by themselves determine either the explanation or the response. Supplementary Figure S3 summarizes this sequence, and the three worked examples in Supplementary Appendix H illustrate it for operational, institutional, and structural profiles.

Equity was the dimension least often addressed in every context examined, and the practical implication is one of disaggregation rather than of a separate indicator set: existing coverage and utilization measures can carry equity information when reported by subgroup, which allows equity to be monitored within the indicator systems health systems already maintain.

Limitations

This study has several limitations, which also help define the next stage of the research agenda taken up in the Conclusion.

First, the synthesis is descriptive: it identifies patterns and associations rather than causal relationships, and the distribution of mitigation strategy categories across weakness categories rests on study-level coding and descriptive cross-tabulation. Because both sets of categories are non-mutually exclusive, these co-occurrence patterns should be read as descriptive associations in the reviewed literature rather than as evidence of comparative effectiveness. No critical appraisal of the included studies was undertaken, so the associations carry no weighting by the robustness of the underlying evidence; a pattern reported mainly in small or weakly designed studies is counted here in the same way as one reported in robust evaluations. This bounds the claims that can be made about the relationship between weakness profiles and responses in particular, and it is the reason those relationships are presented as propositions for contextual investigation rather than as guidance. Because each study was assigned to a single predominant income context and the income-group samples were uneven, the cross-income-group observations are indicative rather than definitive. The multi-country and global category requires particular caution: it is defined by the absence of a single income context rather than by a shared one, and its profile is best read as reflecting the analytical vantage point these studies adopt rather than any single health system. The UMIC group is also small (11 studies reporting a weakness), so observations for that group are particularly tentative.

Second, low-income settings formed the smallest single-tier group in the body of evidence (14 of 128 studies), so the patterns identified may not generalize to highly fragile or conflict-affected contexts. The evidence base may therefore be thinnest precisely where implementation weaknesses are most severe.

Third, no formal appraisal of methodological quality was undertaken, consistent with the interpretive orientation of the synthesis. The findings should accordingly be read as recurrent patterns in the literature rather than as quality-weighted conclusions, and readers assessing the strength of any individual finding should consult the source studies, which are listed with their characteristics in Supplementary Appendix I.

Fourth, the framework is conceptual and has not yet been empirically validated; its value lies in structuring governance decisions rather than in predicting outcomes. Validation could prospectively classify health systems into weakness profiles from baseline indicator signals, record the strategies subsequently deployed, and compare indicator trajectories between systems whose strategies matched their dominant weakness and those where they did not, complemented by a Delphi study on the weighting of indicators within each dimension and by independent replication of the coding.

Together, these limitations define where caution is warranted when applying the framework and motivate the research agenda outlined below.

Conclusions

The evaluation of health policy implementation is fragmented across three literatures that are rarely brought together: what health systems measure, what constrains delivery, and how systems respond. This review set out to integrate them, synthesizing 128 studies through a

788 critical interpretive synthesis in order to propose a framework intended to support governance
789 decisions rather than description alone.

790 In the reviewed literature, evaluation indicators, implementation weaknesses, and mitigation
791 strategies were frequently reported together, although they are usually analysed separately;
792 this suggests a possible mismatch between evaluation systems and the implementation
793 conditions they are intended to monitor. In the studies reviewed, efficiency accounted for the
794 largest share of indicators, equity was the dimension least often addressed in every income
795 group, and sustainability, although widely invoked, was carried by the narrowest set of distinct
796 indicators; implementation weaknesses were rarely reported in isolation, and most studies
797 reporting a weakness described compound profiles.

798 On this basis, the synthesis proposes a framework intended to support the decisions
799 policymakers and managers face. Its main proposed contribution to practice is to allow
800 indicators to be read prospectively: not only as a retrospective record of performance, but as a
801 prompt for examining which weaknesses may be present and which responses could be
802 considered. Three observations run through the framework as a whole: indicators, weaknesses,
803 and strategies were reported together rather than in isolation; context, rather than a universal
804 best-practice list, appears to shape the emphasis of the indicators and strategies reported for a
805 given health system; and the two dimensions beyond service performance were weakly served,
806 though in different ways: equity was the dimension least often addressed at study level, while
807 sustainability, although invoked in most studies, was carried by the narrowest set of distinct
808 recurrent indicators. Both may constitute governance blind spots, but they are blind spots of
809 different kinds. The integrated framework makes these three explicit.

810 The central implication is that health policy evaluation could do more than measure outcomes:
811 it could help systems examine what may be failing, why, and how implementation might be
812 adapted. Proposing that evaluation be repositioned from a retrospective accountability exercise

to a prospective governance instrument is the framework's main contribution; whether this improves implementation decisions remains to be tested. Validating it empirically, through longitudinal and hybrid implementation-effectiveness designs, shared equity and sustainability standards, and investment in health information infrastructure in LIC and LMIC settings, is the natural next step. Even before empirical validation, its value is analytical: a governance-oriented structure intended to support context-sensitive implementation decisions.

List of abbreviations

AMSTAR 2: A MeaSurement Tool to Assess systematic Reviews, version 2

CFIR: Consolidated Framework for Implementation Research

CFIR-ERIC: Consolidated Framework for Implementation Research – Expert Recommendations for Implementing Change

CIS: Critical interpretive synthesis

EPIS: Exploration, Preparation, Implementation, Sustainment framework

GRADE: Grading of Recommendations, Assessment, Development and Evaluations

HIS: Health information system

LIC: Low-income country

LMIC: Lower-middle-income country

MMAT: Mixed Methods Appraisal Tool

OECD: Organisation for Economic Co-operation and Development

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

UHC: Universal health coverage

UMIC: Upper-middle-income country

WHO: World Health Organization

Declarations

Ethics approval and consent to participate

Not applicable. This study is a synthesis of previously published literature and involved no human participants, human data or animals.

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the Zenodo repository, https://doi.org/10.5281/zenodo.22312815. All other data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is co-financed by Component 5 – Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism of the European Union, framed in the Next Generation EU. The funder had no role in the design of the study, in the collection, analysis or interpretation of data, or in writing the manuscript.

Authors' contributions

AMSM: conceptualization, methodology, investigation, data curation, formal analysis, writing – original draft. AR: conceptualization, supervision, writing – review and editing. JCR:

conceptualization, supervision, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Avan BI, Berhanu D, Mekonnen Y et al. Embedding community-based newborn care in the Ethiopian health system: lessons from a 4-year programme evaluation. Health Policy and Planning. 2021;36:i22–i32.

2. Dyeshana HM, Zuma SM. Contributory factors to partial implementation of prevention of mother to child transmission policy in South Africa. BMC Pregnancy and Childbirth 25. 2025.

3. Zurynski Y, Ludlow K, Testa L et al. Built to last? Barriers and facilitators of healthcare program sustainability: a systematic integrative review. Implementation Science. 2023;18:62.

4. Alfvén T, Erkkola T, Ghys PD et al. Global AIDS Reporting-2001 to 2015: Lessons for Monitoring the Sustainable Development Goals. AIDS and Behavior. 2017;21:5–14.

5. Sun D, Ahn H, Lievens T et al. Evaluation of the performance of national health systems in 173 countries (2004-2011). PLOS ONE. 2017;12:e0173346.

6. Thomas JC, Silvestre E, Salentine S et al. What systems are essential to achieving the sustainable development goals and what will it take to marshal them? Health Policy and Planning. 2016;31:1445–1447.

7. Ayowole DJ, Adebajo GO, Lasisi TO et al. Effects of climate change on vaccine storage and cold chain logistics: a qualitative study in Ogun State, Nigeria. BMJ Global Health 10. 2025.

8. Chiliza J, Laing R, Feeley FG et al. Program sustainability post PEPFAR direct service support in the Western Cape, South Africa. PLOS ONE. 2021;16:e0251230.

9. Shang C, Yadav A, Stoklosa M et al. Country-specific costs of implementing the WHO FCTC tobacco control policies and potential financing sources. PLOS ONE. 2018;13:e0204903.

10. KirungaTashobya C, Ssengooba F, Nabyonga-Orem J et al. A critique of the Uganda district league table using a normative health system performance assessment framework. BMC Health Services Research. 2018;18:355.

11. Uzochukwu B, Onwujekwe E, Mbachu C et al. Accountability mechanisms for implementing a health financing option: the case of the basic health care provision fund (BHCPF) in Nigeria. International Journal for Equity in Health. 2018;17:100.

12. Manzi F, Hamon JK, Agbodjavou MK et al. How, why, and under what circumstances can supportive supervision programs improve malaria case management? A realist program theory. Health Policy and Planning. 2025;40:600–612.

13. Oluwole A, Dean L, Lar L et al. Optimising the performance of frontline implementers engaged in the NTD programme in Nigeria: lessons for strengthening community health systems for universal health coverage. Human Resources for Health. 2019;17:79.

14. Engineer CY, Dale E, Agarwal A et al. Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. International Journal of Epidemiology. 2016;45:451–9.

15. Lannes L, Meessen B, Soucat A et al. Can performance-based financing help reach the poor with maternal and child health services? Evidence from rural Rwanda. International Journal of Health Planning and Management. 2016;31:309–348.

16. Jafari M, Habibirad A, Pourtaleb A et al. Health system organizational reform in governing Iranian public hospitals: A content analysis to comprehend the barriers in Board of Trustees' hospitals. International Journal of Health Planning and Management. 2018;33:e612–e620.

17. Dada S, De Brún A, Chelwa N et al. We move together: a realist evaluation of the safe motherhood action groups in eastern province, Zambia. BMJ Global Health 10. 2025.

18. Gai Tobe R, Islam MT, Yoshimura Y et al. Strengthening the community support group to improve maternal and neonatal health seeking behaviors: A cluster-randomized controlled trial in Satkhira District, Bangladesh. PLOS ONE. 2019;14:e0212847.

19. Lemma S, Janson A, Persson LÅ et al. Improving quality and use of routine health information system data in low- and middle-income countries: a scoping review. PLOS ONE. 2020;15:e0239683.

20. Phalkey RK, Yamamoto S, Awate P et al. Challenges with the implementation of an Integrated Disease Surveillance and Response (IDSR) system: systematic review of the lessons learned. Health Policy and Planning. 2015;30:131–43.

21. Danforth K, Ahmad AM, Blanchet K et al. Monitoring and evaluating the implementation of essential packages of health services. BMJ Global Health 8. 2023.

22. Nevedal AL, Opra Widerquist MA, Reardon CM et al. Understanding pathways from implementation to sustainment: a longitudinal, mixed methods analysis of promising practices implemented in the Veterans Health Administration. Implementation Science. 2024;19:34.

23. Waltz TJ, Powell BJ, Fernandez ME et al. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implementation Science. 2019;14:42.

24. Aarons GA, Hurlburt M, Horwitz SM. Advancing a conceptual model of evidence-based practice implementation in public service sectors. Administration and Policy in Mental Health and Mental Health Services Research. 2011;38:4–23.

926 25. Damschroder LJ, Aron DC, Keith RE et al. Fostering implementation of health services
927 research findings into practice: a consolidated framework for advancing implementation
928 science. Implementation Science. 2009;4:50.

929 26. Nambiar D, Sankar DH, Negi J et al. Monitoring Universal Health Coverage reforms in primary
930 health care facilities: Creating a framework, selecting and field-testing indicators in Kerala,
931 India. PLOS ONE. 2020;15:e0236169.

932 27. World Health Organization. Monitoring the Building Blocks of Health Systems: A Handbook
933 of Indicators and Their Measurement Strategies. WHO, Geneva. 2010.
934 https://iris.who.int/handle/10665/258734.

935 28. Smith JD, Li DH, Rafferty MR. The Implementation Research Logic Model: a method for
936 planning, executing, reporting, and synthesizing implementation projects. Implementation
937 Science. 2020;15:84.

938 29. Dixon-Woods M, Cavers D, Agarwal S et al. Conducting a critical interpretive synthesis of the
939 literature on access to healthcare by vulnerable groups. BMC Medical Research
940 Methodology. 2006;6:35.

941 30. Kuchenmüller T, Boeira L, Oliver S et al. Domains and processes for institutionalizing
942 evidence-informed health policy-making: a critical interpretive synthesis. Health Research
943 Policy and Systems. 2022;20:27.

944 31. Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline
945 for reporting systematic reviews. BMJ. 2021;372:n71.

946 32. Hosseinpoor AR, Bergen N, Schlotheuber A et al. Measuring health inequalities in the context
947 of sustainable development goals. Bulletin of the World Health Organization. 2018;96:654–
948 658.

949 33. Wagstaff A, Dmytraczenko T, Almeida G et al. Assessing latent inequality in health systems
950 using SDG health targets: a systematic assessment. The Lancet Global Health. 2018;6:e636–
951 e647.

952 34. Adeleke SA, Ongtengco A, Youssef C et al. Addressing critical barriers for sustainability of
953 asthma stock inhaler policy implementation and resultant programming. Annals of Allergy,
954 Asthma & Immunology. 2024;133:413–421.

955 35. Almeda N, Diaz-Milanes D, Killaspy H et al. Modelling the impact on a local mental health
956 system of previously implemented care programs: the experience of assertive outreach
957 teams in Bizkaia (Spain). Epidemiology and Psychiatric Sciences. 2025;34:e20.

958 36. Amin ET, Omeichu AA, Shu DM et al. Control of antimicrobial resistance in Cameroon:
959 Feasibility of implementing the National Action Plan. Tropical Medicine & International
960 Health. 2021;26:1231–1239.

961 37. Ananthapavan J, Sacks G, Brown V et al. Priority-setting for obesity prevention-The Assessing
962 Cost-Effectiveness of obesity prevention policies in Australia (ACE-Obesity Policy) study.
963 PLOS ONE. 2020;15:e0234804.

964 38. Ansu-Mensah M, Mohammed T, Udoh RH et al. Mapping evidence of free maternal
965 healthcare financing and quality of care in sub-Saharan Africa: a systematic scoping review
966 protocol. Health Research Policy and Systems. 2019;17:93.

967 39. Asefa A, Morgan A, Bohren MA et al. Lessons learned through respectful maternity care
968 training and its implementation in Ethiopia: an interventional mixed methods study.
969 Reproductive Health. 2020;17:103.

970 40. Barasche-Berdah D, Wetzler S, Greenshtein I et al. Municipal community centers as healthy
971 settings: evaluation of a real-world health promotion intervention in Jerusalem. BMC Public
972 Health. 2022;22:1870.

973 41. Bastero P, Staveski SL, Zheleva B et al. Partnership models for the establishment of
974 sustainable paediatric cardiac surgical and cardiac intensive care programmes in low- and
975 middle-income countries. Cardiology in the Young. 2017;27:S55–s60.

976 42. Bawo L, Leonard KL, Mohammed R. Protocol for the evaluation of a quality-based pay for
977 performance scheme in Liberia. Implementation Science. 2015;10:9.

978 43. Birdthistle I, Schaffnit SB, Kwaro D et al. Evaluating the impact of the DREAMS partnership
979 to reduce HIV incidence among adolescent girls and young women in four settings: a study
980 protocol. BMC Public Health. 2018;18:912.

981 44. Bourke CJ, Marrie H, Marrie A. Transforming institutional racism at an Australian hospital.
982 Australian Health Review. 2019;43:611–618.

983 45. Bravata DM, Myers LJ, Homoya B et al. The protocol-guided rapid evaluation of veterans
984 experiencing new transient neurological symptoms (PREVENT) quality improvement
985 program: rationale and methods. BMC Neurology. 2019;19:294.

986 46. Burns A, Kampman H, Magee L et al. Implementation and sustainability of systems change
987 for mental health promotion and substance misuse prevention: a qualitative study. BMC
988 Health Services Research. 2024;24:1477.

989 47. Chai P, Zhang Y, Zhou M et al. Health system productivity in China: a comparison of pre- and
990 post-2009 healthcare reform. Health Policy and Planning. 2020;35:257–266.

991 48. Chantler T, Bell S, Saliba V et al. Is partnership the answer? Delivering the national
992 immunisation programme in the new English health system: a mixed methods study. BMC
993 Public Health. 2019;19:83.

994 49. Chelagat T, Kokwaro G, Odhiambo Onyango J et al. Sustainability drivers and inhibitors for
995 the health system performance improvement projects in selected health facilities in Kenya:
996 a qualitative study. BMJ Open 11. 2021.

997 50. Chien E, Phiri K, Schooley A et al. Successes and Challenges of HIV Mentoring in Malawi: The
998 Mentee Perspective. PLOS ONE. 2016;11:e0158258.

999 51. Christian H, McLaughlin M, Nathan A et al. Scalability and scaling-up strategy of a physical
1000 activity policy intervention in Australian childcare centres. Health Promotion International
1001 40. 2025.

1002 52. Collins D, Diallo BI, Bah MB et al. Evaluation of the first two Frontline cohorts of the field
1003 epidemiology training program in Guinea, West Africa. Human Resources for Health.
1004 2022;20:40.

1005 53. Curfman A, McSwain SD, Chuo J et al. Pediatric Telehealth in the COVID-19 Pandemic Era and
1006 Beyond. Pediatrics 148. 2021.

1007 54. Dayao JKO, Duffy CEL, Cristiano AM et al. Implementation and evaluation of Exercise is
1008 Medicine in primary care clinics within a large academic health system. Family Medicine and
1009 Community Health. 2024;12:e002608.

1010 55. De Neve JW, Boudreaux C, Gill R et al. Harmonizing community-based health worker
1011 programs for HIV: a narrative review and analytic framework. Human Resources for Health.
1012 2017;15:45.

1013 56. Dougherty L, Adediran M, Akinola A et al. An evaluation of a multi-partner approach to
1014 increase routine immunization coverage in six northern Nigerian States. BMC Health
1015 Services Research. 2024;24:951.

1016 57. Duke T, Hwaihwanje I, Kaupa M et al. Solar powered oxygen systems in remote health
1017 centers in Papua New Guinea: a large scale implementation effectiveness trial. Journal of
1018 Global Health. 2017;7:010411.

1019 58. Dullet NW, Geraghty EM, Kaufman T et al. Impact of a University-Based Outpatient
1020 Telemedicine Program on Time Savings, Travel Costs, and Environmental Pollutants. Value
1021 in Health. 2017;20:542–546.

1022 59. Durbin A, Root A, Lawrence HP et al. Building effective public dental care programs: The
1023 critical role of implementation evaluation. Journal of Public Health Dentistry. 2023;83:222–
1024 226.

1025 60. El Bcheraoui C, Kamath AM, Dansereau E et al. Results-based aid with lasting effects:
1026 sustainability in the Salud Mesoamérica Initiative. Globalization and Health. 2018;14:97.

1027 61. Erasmus E, Gilson L, Govender V et al. Organisational culture and trust as influences over the
1028 implementation of equity-oriented policy in two South African case study hospitals.
1029 International Journal for Equity in Health. 2017;16:164.

1030 62. Feldacker C, Makunike-Chikwinya B, Holec M et al. Implementing voluntary medical male
1031 circumcision using an innovative, integrated, health systems approach: experiences from 21
1032 districts in Zimbabwe. Global Health Action. 2018;11:1414997.

1033 63. Foo C, Shrestha P, Wang L et al. Integrating tuberculosis and noncommunicable diseases care
1034 in low- and middle-income countries (LMICs): A systematic review. PLOS Medicine.
1035 2022;19:e1003899.

1036 64. Forero R, Man N, Nahidi S et al. When a health policy cuts both ways: Impact of the National
1037 Emergency Access Target policy on staff and emergency department performance.
1038 Emergency Medicine Australasia. 2020;32:228–239.

1039 65. Fortuna KL, Lebby S, Geiger P et al. Lived Experience-Led Research Agenda to Address Early
1040 Death in People With a Diagnosis of a Serious Mental Illness: A Consensus Statement. JAMA
1041 Network Open. 2023;6:e2315479.

66. Fulcher I, Hedt K, Marealle S et al. Errors in estimated gestational ages reduce the likelihood of health facility deliveries: results from an observational cohort study in Zanzibar. BMC Health Services Research. 2020;20:50.

67. Ganaba R, Ilboudo PGC, Cresswell JA et al. The obstetric care subsidy policy in Burkina Faso: what are the effects after five years of implementation? Findings of a complex evaluation. BMC Pregnancy and Childbirth. 2016;16:84.

68. Greenberg-Worisek AJ, Shippee ND, Schaffhausen C et al. The Learning Health System Competency Appraisal Inventory (LHS-CAI): a novel tool for assessing LHS-focused education needs. Learning Health Systems. 2021;5:e10218.

69. Ha HJ, Han KT, Kim SJ et al. Changes in Saturday outpatient volume and billings after introducing the Saturday incentive programme to clinics in South Korea: a longitudinal cohort study using claims data from 2012 to 2014. BMJ Open. 2016;6:e011248.

70. Hage J, Valadez JJ. Institutionalizing and sustaining social change in health systems: the case of Uganda. Health Policy and Planning. 2017;32:1248–1255.

71. Haq Z, Shaikh BT, Tran N et al. System within systems: challenges and opportunities for the Expanded Programme on Immunisation in Pakistan. Health Research Policy and Systems. 2019;17:51.

72. Haskins L, Chiliza J, Barker P et al. Evaluation of the effectiveness of a quality improvement intervention to support integration of maternal, child and HIV care in primary health care facilities in South Africa. BMC Public Health. 2020;20:318.

73. Henry EG, Thea DM, Hamer DH et al. The impact of a multi-level maternal health programme on facility delivery and capacity for emergency obstetric care in Zambia. Global Public Health. 2018;13:1481–1494.

74. Hobday K, Zwi AB, Homer C et al. Misoprostol for the prevention of post-partum haemorrhage in Mozambique: an analysis of the interface between human rights, maternal health and development. BMC International Health and Human Rights. 2020;20:9.

75. Jaladanki S, Schechter SB, Genies MC et al. Strategies for sustaining high-quality pediatric asthma care in community hospitals. Health Services Research. 2022;57:125–136.

76. Ji S, Lee M, Choi M et al. The impact of the expanded health insurance coverage policy on healthcare spending: evidence from Korea. International Journal for Equity in Health. 2024;23:126.

77. Jones B, Mitchell A, Haynes E et al. Co-designing the implementation of a rural health systems-strengthening rheumatic heart disease program with remote First Nations Australian communities using Theory of Change. BMC Health Services Research. 2025;25:252.

78. Kamya C, Shearer J, Asiimwe G et al. Evaluating Global Health Partnerships: A Case Study of a Gavi HPV Vaccine Application Process in Uganda. International Journal of Health Policy and Management. 2017;6:327–338.

79. Kapologwe NA, Kalolo A, Kibusi SM et al. Understanding the implementation of Direct Health Facility Financing and its effect on health system performance in Tanzania: a non-controlled before and after mixed method study protocol. Health Research Policy and Systems. 2019;17:11.

80. Kassai R, van Weel C, Flegg K et al. Priorities for primary health care policy implementation: recommendations from the combined experience of six countries in the Asia-Pacific. Australian Journal of Primary Health. 2020;26:351–357.

81. Kharaishvili N, Blake JW, Gorsline DH et al. Global Health Security Capacity and Capability Measurement Framework Within the Biological Threat Reduction Program. Health Security. 2021;19:163–172.

82. Kintziger KW, Stone KW, Jagger MA et al. The impact of the COVID-19 response on the provision of other public health services in the U.S.: A cross sectional study. PLOS ONE. 2021;16:e0255844.

83. Kuhlmann E, Falkenbach M, Brînzac MG et al. Tackling the primary healthcare workforce crisis: time to talk about health systems and governance-a comparative assessment of nine countries in the WHO European region. Human Resources for Health. 2024;22:83.

84. Lai J, Klag M, Shikako-Thomas K. Designing a program evaluation for a medical-dental service for adults with autism and intellectual disabilities using the RE-AIM framework. Learning Health Systems. 2019;3:e10192.

85. Lake R, Georgiou A, Li J et al. The quality, safety and governance of telephone triage and advice services - an overview of evidence from systematic reviews. BMC Health Services Research. 2017;17:614.

86. Lamovšek N, Klun M. Evaluation of Biomedical Laboratory Performance Optimisation using the DEA Method. Slovenian Journal of Public Health. 2020;59:172–179.

87. Langlois EV, Mancuso A, Elias V et al. Embedding implementation research to enhance health policy and systems: a multi-country analysis from ten settings in Latin America and the Caribbean. Health Research Policy and Systems. 2019;17:85.

88. Lee RE, Soltero EG, Ledoux TA et al. Sustainability via Active Garden Education: Translating Policy to Practice in Early Care and Education. Journal of School Health. 2019;89:257–266.

89. MacGregor H, McKenzie A, Jacobs T et al. Scaling up ART adherence clubs in the public sector health system in the Western Cape, South Africa: a study of the institutionalisation of a pilot innovation. Globalization and Health. 2018;14:40.

90. Marchal B, Giralt AN, Sulaberidze L et al. Designing and evaluating provider results-based financing for tuberculosis care in Georgia: a realist evaluation protocol. BMJ Open. 2019;9:e030257.

91. Mash RJ, Schouw D. The sustainability of group empowerment and training for people with diabetes in South Africa. South African Family Practice. 2024;66:e1–e11.

92. McDonald T. Lethal ageism in the shadow of pandemic response tactics. International Nursing Review. 2022;69:249–254.

93. McEwen SE, Donald M, Jutzi K et al. Implementing a function-based cognitive strategy intervention within inter-professional stroke rehabilitation teams: Changes in provider knowledge, self-efficacy and practice. PLOS ONE. 2019;14:e0212988.

94. Mirzoev T, Etiaba E, Ebenso B et al. Study protocol: realist evaluation of effectiveness and sustainability of a community health workers programme in improving maternal and child health in Nigeria. Implementation Science. 2016;11:83.

95. Morgan R, Kalbarczyk A, Decker M et al. Gender-responsive monitoring and evaluation for health systems. Health Policy and Planning. 2024;39:1000–1005.

96. Nazar GP, Chang KC, Srivastava S et al. Impact of India's National Tobacco Control Programme on bidi and cigarette consumption: a difference-in-differences analysis. Tobacco Control. 2020;29:103–110.

97. Nelson G, Caplan R, MacLeod T et al. What Happens After the Demonstration Phase? The Sustainability of Canada's At Home/Chez Soi Housing First Programs for Homeless Persons with Mental Illness. American Journal of Community Psychology. 2017;59:144–157.

98. Ngidi WH, Naidoo JR, Ncama BP et al. Mapping evidence of interventions and strategies to bridge the gap in the implementation of the prevention of mother-to-child transmission of

1135 HIV programme policy in sub-Saharan countries: A scoping review. African Journal of
1136 Primary Health Care & Family Medicine. 2017;9:e1–e10.

1137 99. Nishimura F, Carrara AF, Freitas CE. Effect of the Melhor em Casa program on hospital costs.
1138 Revista de Saúde Pública. 2019;53:104.

1139 100. Nungo S, Filippon J, Russo G. Social Health Insurance for Universal Health Coverage in Low
1140 and Middle-Income Countries (LMICs): a retrospective policy analysis of attainments,
1141 setbacks and equity implications of Kenya's social health insurance model. BMJ Open.
1142 2024;14:e085903.

1143 101. Park HJ, Lee JH, Park S et al. Trends in the utilization of dental outpatient services affected
1144 by the expansion of health care benefits in South Korea to include scaling: a 6-year
1145 interrupted time-series study. Journal of Periodontal & Implant Science. 2018;48:3–11.

1146 102. Patel S, Awoonor-Williams JK, Asuru R et al. Benefits and limitations of a community-
1147 engaged emergency referral system in northern Ghana. Global Health: Science and Practice.
1148 2016;4:552–567.

1149 103. Pestana JM. A pioneering healthcare model applying large-scale production concepts:
1150 Principles and performance after more than 11,000 transplants at Hospital do Rim. Revista
1151 da Associação Médica Brasileira. 2016;62:664–671.

1152 104. Pittalis C, Brugha R, Gajewski J. Surgical referral systems in low- And middle-income
1153 countries: A review of the evidence. PLOS ONE 14. 2019.

1154 105. Probandari A, Widjanarko B, Mahendradhata Y et al. The path to impact of operational
1155 research on tuberculosis control policies and practices in Indonesia. Global Health Action.
1156 2016;9:29866.

106. Rahman A, Reza AAS, Bhuiyan BA et al. Equity and determinants of routine child immunisation programme among tribal and non-tribal populations in rural Tangail subdistrict, Bangladesh: a cohort study. BMJ Open. 2018;8:e022634.

107. Rajaeieh G, Bakhtiari A, Gholami M et al. Stakeholders influence to reduce vitamin A deficiency in children aged 15 to 23 months in Iran: an analysis of stakeholders and networking. BMC Nutrition. 2025;11:127.

108. Rashidian A, Jahanmehr N, Farzadfar F et al. Performance evaluation and ranking of regional primary health care and public health Systems in Iran. BMC Health Services Research. 2021;21:1168.

109. Requejo J, Moran AC, Monet JP. Accountability for maternal and newborn health: why broader health system determinants matter. PLOS ONE. 2024;19:e0300429.

110. Ruzafa-Martinez M, Harillo-Acevedo D, Ramos-Morcillo AJ. Monitoring of the Implementation of a Breastfeeding Guideline for 6 Years: A Mixed-Methods Study Using an Interrupted Time Series Approach. Journal of Nursing Scholarship. 2021;53:358–368.

111. Saldana L, Dakof G, Cruden G et al. Facilitating implementation of a substance use intervention for youth: Outcomes from a randomized trial of the SIC-coaching implementation strategy. Journal of Substance Use and Addiction Treatment. 2025;173:209696.

112. Saprii L, Richards E, Kokho P et al. Community health workers in rural India: analysing the opportunities and challenges Accredited Social Health Activists (ASHAs) face in realising their multiple roles. Human Resources for Health. 2015;13:95.

113. Sarkies MN, Francis-Auton E, Long JC et al. Implementing large-system, value-based healthcare initiatives: a realist study protocol for seven natural experiments. BMJ Open. 2020;10:e044049.

114. Seppey M, Ridde V, Touré L et al. Sustainability assessment of a donor-funded results-based financing pilot in Mali. Globalization and Health. 2017;13:86.

115. Shams L, Zamani Fard M, Nasiri T et al. Community health workers (Behvarz) in primary health care: a qualitative inductive content analysis of challenges. Australian Journal of Primary Health. 2023;29:428–436.

116. Shigayeva A, Coker RJ. Communicable disease control programmes and health systems: an analytical approach to sustainability. Health Policy and Planning. 2015;30:368–85.

117. Shumayla S, Lalchandani K, Verma D et al. Sustainability of evidence-based policy engagement model: A case study of Advance Family Planning initiative in India. PLOS ONE. 2025;20:e0320295.

118. Smith N, Mitton C, Hiltz MA et al. A Qualitative Evaluation of Program Budgeting and Marginal Analysis in a Canadian Pediatric Tertiary Care Institution. Applied Health Economics and Health Policy. 2016;14:559–68.

119. Soban LM, Kim L, Yuan AH et al. Organisational strategies to implement hospital pressure ulcer prevention programmes: findings from a national survey. Journal of Nursing Management. 2017;25:457–467.

120. Soi C, Shearer J, Chilundo B et al. Global health systems partnerships: a mixed methods analysis of Mozambique's HPV vaccine delivery network actors. BMC Public Health. 2020;20:862.

121. Sow M, De Spiegelaere M, Raynault MF. Evaluating the effect of income support policies on social health inequalities (SHIs) at birth in Montreal and Brussels using a contextualised comparative approach and model family method: a study protocol. BMJ Open. 2018;8:e024015.

122. Stuikyte R, Varentsov I, Malikov N et al. Sustainability of opioid agonist therapy programmes during transition from global fund support. Harm Reduction Journal. 2024;21:183.

123. Sun R, Mendez D. Finding the optimal mix of smoking initiation and cessation interventions to reduce smoking prevalence. PLOS ONE. 2019;14:e0212838.

124. Tremblay St-Germain A, Devitt KS, Kagedan DJ et al. The impact of a clinical pathway on patient postoperative recovery following pancreaticoduodenectomy. HPB. 2017;19:799–807.

125. Uneke CJ, Sombie I, Uro-Chukwu HC et al. Using equitable impact sensitive tool (EQUIST) to promote implementation of evidence informed policymaking to improve maternal and child health outcomes: a focus on six West African Countries. Globalization and Health. 2018;14:104.

126. Vale DB, Menin TL, Bragança JF et al. Estimating the public health impact of a national guideline on cervical cancer screening: an audit study of a program in Campinas, Brazil. BMC Public Health. 2019;19:1492.

127. Ventura M, Fusco D, Bontempi K et al. Regional Outcome Evaluation Program (P.Re.Val.E.): Reduction of inequality in access to effective health care in the Lazio region of Italy (2012-2015). PLOS ONE. 2018;13:e0194972.

128. Wang J, Bai L, Xu X. Disparities in awareness and utilisation of national essential public health services in China. BMJ Open. 2024;14:e080756.

129. Watkins DA, Olson ZD, Verguet S et al. Cardiovascular disease and impoverishment averted due to a salt reduction policy in South Africa: an extended cost-effectiveness analysis. Health Policy and Planning. 2016;31:75–82.

130. Wolk CB, Last BS, Livesey C et al. Addressing Common Challenges in the Implementation of Collaborative Care for Mental Health: The Penn Integrated Care Program. Annals of Family Medicine. 2021;19:148–156.

131. Wrona J, Hardy P, Youssef C et al. Stock Inhalers: A Qualitative Data Analysis of Illinois Health Policy Trials and Triumphs. Journal of School Health. 2024;94:918–928.

132. Xie W, Shi L, Liu M et al. Disparities and effectiveness of COVID-19 vaccine policies in three representative European countries. International Journal for Equity in Health. 2024;23:16.

133. Yazdizadeh B, Mohtasham F, Velayati A. Impact assessment of Iran's health technology assessment programme. Health Research Policy and Systems. 2018;16:15.

134. Yee LM, Miller ES, Statton A et al. Sustainability of Statewide Rapid HIV Testing in Labor and Delivery. AIDS and Behavior. 2018;22:538–544.

135. Yoshioka K, Tercero D, Pérez B et al. Implementing a vector surveillance-response system for chagas disease control: a 4-year field trial in Nicaragua. Infectious Diseases of Poverty. 2017;6:18.

136. Zakumumpa H, Bennett S, Ssengooba F. Accounting for variations in ART program sustainability outcomes in health facilities in Uganda: a comparative case study analysis. BMC Health Services Research. 2016;16:584.

137. Zhang L, Wang X, Yao H et al. Impact of national drug price negotiation policy on the accessibility and utilization of PCSK9 inhibitors in China: an interrupted time series analysis. International Journal for Equity in Health. 2024;23:116.

138. Zhang X, Tan SS, Fierloos I et al. Evaluation design of the Social Engagement Framework for Addressing the Chronic-disease-challenge (SEFAC): a mindfulness-based intervention to promote the self-management of chronic conditions and a healthy lifestyle. BMC Public Health. 2019;19:664.

1250 139. Zhu T, Chen C, Zhang X et al. Differences in inpatient performance of public general
1251 hospitals following implementation of a points-counting payment based on diagnosis-
1252 related group: a robust multiple interrupted time series study in Wenzhou, China. BMJ
1253 Open. 2024;14:e073913.

1254 140. Gathara D, Zosi M, Serem G et al. Developing metrics for nursing quality of care for low-
1255 and middle-income countries: a scoping review linked to stakeholder engagement. Human
1256 Resources for Health. 2020;18:34.

1257 141. Karimi Jashni Y, Emari F, Morris M et al. Indicators of integrating oral health care within
1258 universal health coverage and general health care in low-, middle-, and high-income
1259 countries: a scoping review. BMC Oral Health. 2023;23:251.

1260 142. Victora CG, Requejo JH, Barros AJD et al. Countdown to 2030 for reproductive, maternal,
1261 newborn, child, and adolescent health and nutrition. The Lancet Global Health.
1262 2018;6:e1196–e1252.

1263 143. Kok MC, Vallières F, Tulloch O et al. Does supportive supervision enhance community health
1264 worker motivation? A mixed-methods study in four African countries. Health Policy and
1265 Planning. 2018;33:988–998.

1266 144. Schloemer T, Schröder-Bäck P. Criteria for evaluating transferability of health interventions:
1267 a systematic review and thematic synthesis. Implementation Science. 2018;13:88.

1268 145. Bullock HL, Lavis JN, Wilson MG et al. Understanding the implementation of evidence-
1269 informed policies and practices from a policy perspective: a critical interpretive synthesis.
1270 Implementation Science. 2021;16:18.

1271 146. Varallyay NI, Langlois EV, Tran N et al. Health system decision-makers at the helm of
1272 implementation research: development of a framework to evaluate the processes and
1273 effectiveness of embedded approaches. Health Research Policy and Systems. 2020;18:64.

147. Pacifico Silva H, Lehoux P, Miller FA et al. Introducing responsible innovation in health: a policy-oriented framework. Health Research Policy and Systems. 2018;16:90.

148. Baxter S, Johnson M, Chambers D et al. Towards greater understanding of implementation during systematic reviews of complex healthcare interventions: the framework for implementation transferability applicability reporting (FITAR). BMC Medical Research Methodology. 2019;19:80.

149. Munthe-Kaas H, Booth A, Nøkleby H et al. The transfer approach for assessing the transferability of systematic review findings. BMC Medical Research Methodology. 2020;20:19.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 10SupplementaryMaterial.docx