

Journal of Holistic Health and Wellness

ISSN 3143-6430

DAC Insight Publishers

<https://journals.dacinsightpublishers.com/JHHW>



DESIGNING A QUALITY ASSURANCE FRAMEWORK FOR EVALUATING ADMINISTRATIVE RELIABILITY IN HOME-BASED HEALTHCARE TRANSITIONS

Moriyanu Ogundipe^{1*}, Schilder Ominde²

¹ Keller Graduate School of Management, DeVry University, Chicago, IL, USA

² Independent Researcher, USA

Corresponding Author*: ogundmo@gmail.com

ABSTRACT

Administrative steps such as referral intake, order authentication, medication reconciliation, and structured handoff shape continuity of care when patients enter Medicare-certified home health, yet the indicators used to track these steps are seldom tested before they are used to judge agencies. This paper proposes a quality assurance framework that specifies the evidence an administrative reliability indicator must carry before routine use. Grounded in Donabedian's structure-process-outcome model, argument-based validity theory, and generalizability theory, the framework defines three indicator classes (structural, process-reliability, and criterion) and a cross-cutting equity-differential layer; and gives ten indicators operational definitions with numerators, denominators, time windows, data sources, and coding rules. It sets explicit decision rules: content validity indices, use-tiered interrater agreement thresholds supplemented by prevalence-robust coefficients, agency-level signal-to-noise reliability, and quantified tests of differential reliability across rurality, language, race and ethnicity, dual eligibility, and disability entitlement. Eight testable propositions, a separately stated guiding principle, and a four-stage validation pathway govern certification, with distinct labels for indicators whose reliability rests on documentation alone and those verified against operational records. The framework is compared with existing measure-evaluation standards; its distinct contributions are equity-stratified reliability as a certification gate and the documentation-operational distinction. No indicator has yet been validated empirically; the paper supplies the architecture, decision thresholds, and sample-size guidance for doing so.

Keywords: *Quality assurance; administrative reliability; measurement validation; interrater reliability; home health agencies; care transitions; health equity; Donabedian framework*

Received 4 May 2026; Received in revised form 15 May 2026; Accepted 25 May 2026, Available online 30 May 2026

1. INTRODUCTION

When a patient moves from hospital or post-acute care into home health, continuity depends on a chain of administrative steps: a referral that carries the information the agency needs, eligibility and certification documentation, timely authentication of orders, reconciliation of the medication regimen, a handoff that transfers substantive clinical content, and a first visit and escalation pathway that work when needed. Each step is a documented failure point in care transitions research (Coleman *et al.*, 2006; Kripalani *et al.*, 2007; Naylor *et al.*, 2011). Agencies and multi-agency collaboratives increasingly build their own indicators for these steps, for example within the quality assessment and performance improvement (QAPI) program required under 42 CFR 484.65, and use them to audit staff, target improvement resources, and compare sites (Centers for Medicare & Medicaid Services [CMS], 2026c).

Such indicators are usually computed from chart review, and chart review introduces error from abstractors, from audit periods, and from differences in documentation practice between agencies (Allison *et al.*, 2000; Vassar & Holzmman, 2013). Mature standards already exist for evaluating measures: the National Quality Forum (NQF) criteria and the guidebook of the current consensus-based entity (Battelle, 2024; National Quality Forum, 2021), the CMS Measures Management System Blueprint (CMS, 2026a), the COSMIN taxonomy (Mokkink *et al.*, 2010), the Standards for Educational and Psychological Testing (American Educational Research Association *et al.*, 2014), and argument-based validation (Kane, 2013). Publicly reported Home Health Quality Reporting Program measures pass through these processes and are accompanied by published specifications and risk-adjustment documentation (CMS, 2026e). Locally built administrative process indicators typically do not, and the standards themselves were written for nationally specified measures or patient-reported instruments rather than for chart-abstracted indicators in a setting where payment rules strongly shape what is documented.

This paper proposes a quality assurance framework for that gap. It asks four questions: what validity evidence is relevant to an administrative reliability indicator; what reliability evidence chart-abstracted indicators require, at the chart and at the agency level; how criterion validity should be tested when the outcomes involved are themselves imperfectly measured; and what staged pathway allows a research team or collaborative to certify an indicator for a stated use. The literature base covers publications to June 2026. The paper reports no empirical validation; its contribution is a measurement specification with explicit decision rules, offered for future empirical testing. Measurement validity, causal effectiveness, and regulatory endorsement are treated as three different claims, and the framework addresses only the first. The authors' related manuscripts address governance and intervention design for the same administrative processes (Ogundipe & Ominde, 2024, 2025); every element of that work on which the present argument relies is restated here, so the framework can be read and applied independently.

2 THEORETICAL GROUNDING

2.1 Administrative reliability as a process construct

Administrative reliability is defined here as the consistency with which the administrative steps required in a care transition are completed correctly, on time, and with substantive content. In Donabedian's terms these are process attributes: they describe what was done rather than what resulted. Donabedian (1988, p. 1745) grounds the structure-process-outcome approach in the premise that good structure increases the likelihood of good process, and good process increases

the likelihood of good outcomes. The framework takes that premise as a hypothesis to be tested for each indicator rather than as a property every process measure inherits. Haggerty *et al.* (2003) distinguish informational, management, and relational continuity; the administrative steps addressed here serve the first two, which fixes the construct each indicator must represent. An indicator of handoff quality, for instance, must capture whether clinically relevant information moved between providers, not merely whether a handoff form exists in the record.

2.2 Validity as an argument about interpretation and use

Cronbach and Meehl (1955) established that construct validity rests on a network of predicted relationships rather than a single correlation, and Messick (1995) added that the consequences of using a measure for decisions are part of its validity. Kane (2013) organizes this evidence as an interpretation and use argument: the user first states what a score is taken to mean and what decision it will support, and the evidence required then scales with the stakes of that decision. The Standards for Educational and Psychological Testing adopt the same position and treat fairness across subgroups as part of validity (American Educational Research Association *et al.*, 2014). The framework applies this logic directly. An indicator used for internal monitoring within one agency makes a narrower claim than the same indicator used to rank agencies or to inform payment, so the reliability thresholds in Section 6 are tiered by intended use. This also explains why equity testing belongs inside validation: an indicator whose error differs across patient groups supports different inferences for those groups.

2.3 Reliability as generalizability across raters, occasions, and sites

A chart-abstracted indicator carries error from at least three facets: the abstractor, the audit occasion, and the agency whose documentation conventions are being read. Generalizability theory partitions observed variance among such facets and estimates how dependable a score is for a given design (Brennan, 2001); the intraclass correlation forms described by Shrout and Fleiss (1979) are special cases of that approach. Two consequences follow. First, agreement between two abstractors on the same charts is necessary but not sufficient, because it leaves occasion and agency facets unexamined. Second, when indicators are used to compare agencies, the relevant quantity is agency-level reliability, the share of variance in agency scores attributable to true differences between agencies rather than sampling and abstraction noise (Adams, 2009). The chart review literature adds a further distinction: abstraction error, where reviewers read the same record differently, is separate from recording error, where the record itself misstates what happened (Allison *et al.*, 2000). Section 6.7 builds on that distinction.

2.4 Scope of the framework

The framework validates administrative process indicators that an agency or research team can construct from records kept under the Medicare Conditions of Participation at 42 CFR Part 484, from the Outcome and Assessment Information Set (OASIS), and from QAPI documentation (CMS, 2023, 2026c, 2026d). It does not address clinical quality measures, patient-reported outcome measures, or financial measures, each of which has its own validation literature. A validated indicator under this framework is not thereby an endorsed or CMS-adopted measure. Where Stages 3 and 4 require claims data or linked utilization data, the requirement is stated explicitly rather than assumed to be available. Every validation claim is time-stamped and setting-stamped: it applies to the population, period, and data definitions under which it was established, and Section 6.8 specifies when it must be refreshed (Kane, 2013).

3 POSITIONING AGAINST EXISTING MEASURE-EVALUATION STANDARDS

3.1 What existing standards already provide

Most components of this framework are established practice, and Table 1 sets the framework beside the five standards most relevant to it. The NQF criteria, now maintained through the consensus-based entity's Endorsement and Maintenance Guidebook, require evidence of importance, reliability and validity testing at the data-element or score level, feasibility, and usability, with periodic maintenance review (Battelle, 2024; National Quality Forum, 2021). The CMS Blueprint governs the measure lifecycle from conceptualization through testing, implementation, and ongoing evaluation (CMS, 2026a). COSMIN supplies a taxonomy of measurement properties, including reliability, measurement error, criterion validity, cross-cultural validity, and responsiveness, developed for health measurement instruments (Mokkink *et al.*, 2010). The Standards (American Educational Research Association *et al.*, 2014) and Kane's (2013) argument-based approach supply the general logic of validity as evidence for a stated interpretation and use.

Table 1: Comparison of the proposed framework with existing measure-evaluation standards

Standard	Primary object	Reliability evidence	Validity evidence	Subgroup equity requirement	or	Re-evaluation
NQF criteria; CBE Endorsement and Maintenance Guidebook	Nationally specified performance measures submitted for endorsement	Data-element or score-level testing	Data-element or score-level testing; systematic face validity in some cases	Disparities social adjustment considered; stratified reliability not required before endorsement	and risk	Scheduled maintenance review
CMS MMS Blueprint	Measures developed for CMS programs	Testing during development	Testing during development and evaluation	Equity considered during development		Annual updates and comprehensive reevaluation
COSMIN taxonomy	Health measurement instruments, mainly patient-reported	Reliability and measurement error as separate properties	Content, structural, hypotheses-testing, criterion validity; responsiveness	Cross-cultural validity and invariance for multi-item instruments		Not specified at instrument level
AERA, APA, NCME Standards (2014)	Educational and psychological tests	Reliability or precision and errors of measurement	Five sources of validity evidence	Fairness chapter; subgroup analysis		When use or population changes

Standard	Primary object	Reliability evidence	Validity evidence	Subgroup equity requirement	or Re-evaluation
Kane (2013) argument-based validation	Any score interpretation and use	Generalization inference	Extrapolation and implication inferences	Through evaluation of consequences	When the stated use changes
Present framework	Chart-abstracted administrative process indicators in home health transitions	Interrater (kappa with PABAK and ICC); agency-level; documentation versus operational	Content (CVI); criterion with multilevel adjustment; convergent with HHVBP; responsiveness	Differential reliability and effect modification across five strata required before certification	Triggered by OASIS or Conditions of Participation definition changes and periodic review

3.2 What the present framework adds

Against this background, the framework claims four specific additions rather than a new theory of measurement. First, it makes equity-stratified reliability a precondition for certification. Existing criteria ask whether disparities should be examined or social risk adjusted, but they do not require that an indicator demonstrate adequate reliability within each relevant patient stratum before use (Battelle, 2024; National Quality Forum, 2021). Second, it separates documentation reliability from operational reliability and attaches different permitted uses to each, a distinction the chart review literature recognizes but measure-evaluation standards do not operationalize (Allison *et al.*, 2000). Third, it sets class-specific gates for chart-abstracted indicators whose prevalence is driven toward ceiling by payment rules, where kappa alone is known to mislead (Feinstein & Cicchetti, 1990). Fourth, it ties re-validation to changes in source definitions, such as OASIS revisions, at the level of an agency or collaborative rather than a national steward. The Blueprint's lifecycle maintenance performs a comparable function for national measures (CMS, 2026a); the framework brings that discipline to indicators that no national steward maintains.

4 LITERATURE BASIS FOR THE FRAMEWORK

4.1 Identification and appraisal of sources

Sources were identified through targeted searches of MEDLINE (via PubMed), CINAHL, Scopus, and the Cochrane Library for records published from January 1970 to June 2026, supplemented by backward citation tracking and by CMS regulatory and measure-specification documents. Search strings combined setting terms (home health, home care, transitional care, care transition), process terms (referral, medication reconciliation, handoff, verbal order, face-to-face, discharge communication), and measurement terms (reliability, agreement, kappa, validity, OASIS, chart review, quality indicator). Sources were included when they reported empirical data on an administrative transition process, its measurement, or its association with outcomes, or when they set out measurement methods applicable to chart-abstracted indicators. Each source was coded for the measurement claim it could inform (construct, criterion, reliability, or equity) and for design.

Conference abstracts, commentaries, and protocols were not given evidential weight, and a protocol, its completed review, and derivative summaries were counted as one body of evidence (Redmond *et al.*, 2018). The review is a structured narrative synthesis, not a systematic review, and Section 9.5 records that limitation (Vassar & Holzmann, 2013).

4.2 Evidence bearing on construct definition

The construct is well specified at the level of mechanism. Haggerty *et al.* (2003) define the continuity types that administrative steps serve, and the Institute of Medicine (2001) argues that outcome reporting without process measurement conceals the mechanism a health system is accountable for. Kripalani *et al.* (2007) document deficits in information transfer at discharge, including absent or delayed discharge summaries, which define the content an intake and handoff indicator must capture. Coleman *et al.* (2006), Jack *et al.* (2009), and Naylor *et al.* (2004) tested interventions built around medication self-management, structured discharge information, and timely follow-up, and in doing so identified the specific process elements that the indicators in Section 5 operationalize. This literature supports content validity: it shows what each indicator should represent. It does not show that any particular operational definition captures that content consistently, which is a reliability question.

4.3 Evidence bearing on criterion relationships

Studies linking transition processes to readmission supply indirect criterion evidence. Pesko *et al.* (2018) associate nurse-physician communication in home health with readmission, and a congestive heart failure literature spanning several decades links the adequacy and timing of home care to readmission (Anderson *et al.*, 2005; Li *et al.*, 2004; Martens, 2000; Proctor *et al.*, 2000). Navarro *et al.* (2012) show that routinely collected data can identify readmission risk among Medicare home health patients, and Kellogg *et al.* (1991) report readmission control through community-based follow-up. Lodge *et al.* (2024), studying children with tracheostomies, and Treece *et al.* (2018), reviewing hospice, examined populations whose care processes differ from those of adult Medicare home health patients; associations observed in adult cohorts therefore cannot be assumed to hold for them, and the framework requires criterion testing within the population of intended use. Readmission is also shaped by factors outside agency control, including discharge planning models (Damiani *et al.*, 2009), hospital safety climate (Hansen *et al.*, 2011), and county-level context (Nuckols, 2015). These studies differ in design and report associations of varying size, so no pooled effect is used here. They justify expecting modest associations and testing them with case-mix adjustment.

4.4 Evidence bearing on reliability

Direct reliability evidence for administrative transition indicators is scarce, but adjacent measurement literature is substantial. Madigan and Fortinsky (2004) examined interrater reliability of OASIS items under field conditions and found that agreement varied considerably across items, and Kinatukara *et al.* (2005) compared several methodological approaches to assessing OASIS reliability and validity, showing that conclusions depend on design choices. The chart abstraction literature specifies the conditions for reliable abstraction: explicit operational definitions, abstractor training, blinding, and ongoing reliability monitoring (Allison *et al.*, 2000; Vassar & Holzmann, 2013). The medication reconciliation literature illustrates the gap. Redmond *et al.* (2018) pooled reductions in medication discrepancies across trials but rated the evidence as very low certainty, and consistent pooled effects say nothing about whether two reviewers classify the same discrepancy the same way. Studies documenting reconciliation gaps (Bartzak, 2022; Becerra, 2019; Heffner *et al.*, 2020; Martin, 2012; Tamiru *et al.*, 2018) apply broadly similar discrepancy

classifications without reporting interrater statistics. The framework therefore requires reliability to be measured directly.

4.5 Evidence bearing on equity

Disparities in home health access and outcomes are documented. Fashaw-Walters *et al.* (2022) found that Black and Hispanic patients were less likely than White patients to receive care from high-quality agencies, and Narayan and Scafide (2017) synthesized racial and ethnic outcome disparities across home health studies. Squires *et al.* (2022) examined patient language preference as a factor in 30-day readmission risk from home health care, and Distelhorst and Hansen (2022) linked neighborhood context to heart failure readmission among older adults receiving transitional care. These findings bear on measurement as well as on care. If documentation is less complete or harder to interpret for patients who need an interpreter, or for agencies serving rural areas, an indicator can be less reliable in exactly the groups where disparities are greatest. OASIS-E introduced standardized items for ethnicity (A1005), race (A1010), and preferred language and interpreter need (A1110), which make stratification feasible from routine data (CMS, 2023). Section 6.5 sets the strata and decision rules accordingly.

4.6 Clinical heterogeneity and summary of the evidence

Home health caseloads are clinically heterogeneous, spanning psychiatric care (Gerace, 1990; Thobaben, 2012), HIV management (Clark, 1991), and older adults moving into home care after hospitalization or nursing home stays (Boling & Parsons, 2007; Heeke *et al.*, 2014; Magilvy & Lakomy, 1992). Umbrella and program-level evaluations of transitional care report variable effects across populations (Hung & Leidig, 2015; Joo & Liu, 2023). For measurement, this heterogeneity implies that reliability and criterion properties may differ by caseload, so pooled estimates should be accompanied by stratified ones. Taken together, the literature supports construct definition and plausible, modest criterion relationships; it documents disparities that make equity testing necessary; and it supplies tested methods for assessing OASIS and chart-abstraction reliability. What it does not supply is reliability, agency-level, or equity-stratified evidence for the specific administrative indicators defined below. The gap is therefore narrower than an absence of measurement literature, but it is real.

5 THE FRAMEWORK: FAILURE MODES, INDICATOR CLASSES, AND OPERATIONAL DEFINITIONS

5.1 Six administrative failure modes

The indicators are organized around six administrative failure modes, summarized in Table 2 with the regulatory anchor for each and the indicators that target it. Each failure mode has a core element, the administrative step and its timing or content standard, which the indicator measures, and adaptable elements, such as which staff role performs the step or which software records it, which vary legitimately between agencies. The distinction matters for responsiveness testing (Proposition 7): an indicator should change when its core element changes, regardless of how the adaptable elements are configured. Regulatory anchors are drawn from the Conditions of Participation and the certification rules for Medicare home health (CMS, 2026c, 2026b). Two structural indicators support all six failure modes by describing agency capacity rather than a single transition step, consistent with the structural layer of Donabedian's (1988) model.

Table 2: Administrative failure modes, regulatory anchors, and targeting indicators

Failure mode	Core element measured	Regulatory anchor	Indicator(s)
F1. Incomplete referral at intake	Required referral elements present when the referral is accepted	Agency intake policy; 42 CFR 484.60 (care planning and coordination)	I3 Referral completeness
F2. Unsupported eligibility and certification	Homebound status, certification elements, and face-to-face encounter documented	42 CFR 424.22 (certification; face-to-face encounter)	I4 Eligibility and homebound documentation; I6 Face-to-face encounter documentation
F3. Delayed authentication of orders	Verbal orders authenticated and dated within the agency's policy window	42 CFR 484.60(b)(4)	I5 Verbal-order authentication timeliness
F4. Unreconciled medication regimen	Discrepancies identified and resolved with the practitioner	42 CFR 484.55(c) (drug regimen review); OASIS M2001 to M2005	I7 Medication reconciliation accuracy
F5. Nominal handoff	Handoff contains situation, background, assessment, and recommendation content	42 CFR 484.60 (coordination); 42 CFR 484.110(a)(6) (transfer and discharge summaries)	I8 SBAR handoff substantive completion
F6. Delayed first contact or unanswered escalation	Initial visit within 48 hours; after-hours calls answered and acted on	42 CFR 484.55(a)(1); agency on-call policy	I9 Initial visit within 48 hours; I10 After-hours escalation response
Capacity (all modes)	Agency-level systems that support the steps above	42 CFR 484.65 (QAPI)	I1 QAPI program maturity; I2 Electronic referral capability

5.2 Indicator classes and the equity-differential layer

Table 3 defines three indicator classes, each with a primary validity requirement, and an equity-differential layer that is an analysis applied to every indicator rather than an indicator in its own right. Structural indicators describe agency capacity and require content validity plus agreement between independent verifiers. Process-reliability indicators describe whether a step occurred as defined and require interrater and agency-level reliability. Criterion indicators are process indicators for which a link to a downstream outcome is claimed; they require the same reliability

evidence plus criterion validity and responsiveness. The classes follow Donabedian's (1988) sequence. Ten indicators result: two structural, four process-reliability, and four criterion. The convergent comparison with the Home Health Value-Based Purchasing (HHVBP) Total Performance Score is a validation analysis (Proposition 4), not an indicator, and is therefore listed in Section 7 rather than in Table 3 (CMS, 2021). Every process-reliability and criterion indicator carries an interrater reliability requirement, including those computed from timestamps, because timestamp extraction also involves abstractor judgment about which record entry counts (Allison *et al.*, 2000).

Table 3: Indicator classes, primary validity requirements, and Donabedian mapping

Class or layer	Indicators	Primary validity requirement	Additional requirement	Donabedian category
Structural	I1, I2	Content validity (I-CVI, S-CVI/Ave)	Inter-verifier agreement on the capacity checklist	Structure
Process-reliability	I3, I4, I5, I6	Interrater reliability	Agency-level reliability when agencies are compared	Process
Criterion	I7, I8, I9, I10	Criterion validity against an independently ascertained outcome	Interrater and agency-level reliability; responsiveness	Process linked to outcome
Equity-differential layer	Applied to I1 to I10	Differential reliability and effect modification across strata	Prespecified margins (Section 6.5)	Cross-cutting

5.3 Operational definitions

Construct validity cannot be assessed for an indicator that is named but not defined. Table 4 gives each indicator's numerator, denominator, unit of analysis, and time window, and Table 5 gives its data source, coding rule, and the minimum validation evidence required. OASIS item identifiers follow the OASIS-E and OASIS-E1 guidance manuals (CMS, 2023, 2024), and regulatory references follow the current Code of Federal Regulations (CMS, 2026c, 2026b). Referral completeness (I3) is not constructed from OASIS; it uses the agency's intake record. The coding rule for SBAR handoff substantive completion (I8) operationalizes the distinction between a handoff that exists and one that transfers content: each of the four SBAR components is scored present only if it contains patient-specific information, and a component that restates a template heading or a diagnosis label alone is scored absent. Face-to-face documentation (I6) is restricted to traditional Medicare fee-for-service episodes, because the certification rule at 42 CFR 424.22 defines the requirement for those episodes; Medicare Advantage organizations apply their own authorization processes, so the same documentation standard cannot be assumed to apply. Agencies adopting the framework should record the version of each definition used, as Section 6.8 requires (Vassar & Holzmann, 2013).

Table 4: Operational definitions of the ten indicators

Indicator	Numerator	Denominator	Unit; time window
I1 QAPI program maturity	Sum of rubric levels achieved across eight QAPI domains (0 to 3 each)	Not applicable (scored rubric, range 0 to 24)	Agency; annual
I2 Electronic referral capability	Checklist capabilities verified present (of six)	Six specified capabilities	Agency; annual
I3 Referral completeness	Accepted referrals with all eight required elements present at acceptance	Referrals accepted in period	Referral; monthly
I4 Eligibility and homebound documentation	Episodes with homebound criteria and certification elements documented before the first claim	Episodes started in period	Episode; monthly
I5 Verbal-order authentication timeliness	Verbal orders authenticated and dated within the agency policy window	Verbal orders received in period	Order; monthly
I6 Face-to-face encounter documentation	FFS episodes with a qualifying encounter documented within 90 days before or 30 days after start of care	Traditional Medicare FFS episodes started in period	Episode; monthly
I7 Medication reconciliation accuracy	Episodes in which all reference-standard discrepancies were identified and resolved with the practitioner	Episodes with a drug regimen review in period	Episode; monthly
I8 SBAR handoff substantive completion	Handoffs with all four SBAR components scored substantive	Handoffs at start of care, transfer, and discharge in period	Handoff; monthly
I9 Initial visit within 48 hours	Episodes with initial visit within 48 hours of referral, return home, or ordered start date	Episodes started in period	Episode; monthly
I10 After-hours escalation response	After-hours calls answered by clinician and acted on within 30 minutes	After-hours patient calls in period	Call; monthly

Table 5: Data sources, coding rules, and minimum validation evidence

Indicator	Data source	Coding rule	Minimum evidence for certification
I1	QAPI plan, minutes, and project records (42 CFR 484.65)	Rubric anchors define each level; level credited only with documentary evidence	I-CVI and S-CVI/Ave; inter-verifier ICC; internal structure examined only if used as a multi-item scale
I2	System configuration and audit logs	Capability credited only when demonstrated in a live record	I-CVI; inter-verifier kappa per checklist item
I3	Intake record and referral packet (not OASIS)	Eight elements: demographics, payer verification, referring practitioner, diagnoses, medication list, recent discharge summary, service orders, certifying practitioner	Kappa with PABAK and AC1; agency-level reliability if compared
I4	Certification and plan of care; visit notes (42 CFR 424.22)	Homebound criteria met only if both criteria are supported by documented findings	Kappa with PABAK and AC1
I5	Order log and e-signature audit trail (42 CFR 484.60(b)(4))	Authentication date taken from signature record, not order entry date	Kappa with PABAK and AC1; operational verification against audit trail
I6	Certification documents; encounter note (42 CFR 424.22(a)(1)(v))	Encounter qualifies if performed by a permitted practitioner, in person or by permitted telehealth, related to primary reason for care	Kappa with PABAK and AC1
I7	Drug regimen review; OASIS M2001, M2003, M2005, N0415; reference medication list	Reference standard set by a pharmacist reviewer blind to agency coding	Kappa with PABAK and AC1; criterion against independently ascertained medication-related ED visit or hospitalization
I8	Transfer, discharge, and start-of-care communication notes	Component substantive only if patient-specific; template text scored absent	Kappa with PABAK and AC1; criterion against plan-of-care initiation delay

Indicator	Data source	Coding rule	Minimum evidence for certification
I9	OASIS M0030, M0102, M0104; visit timestamps	Clock starts at referral, return home, or ordered start date, whichever applies	ICC(2,1) or kappa on the binary rule; test-retest; criterion against 30-day unplanned hospitalization
I10	Telephony and on-call logs	Response time from call log, action from clinical note	ICC(2,1); operational verification against telephony record; criterion against ED use within 72 hours

5.4 Why a single validity coefficient is not sufficient

Reporting one association between an indicator and an outcome, and calling the indicator validated, conflates evidence that should be kept separate (Cronbach & Meehl, 1955). An indicator needs, at minimum, evidence of reliability, evidence of responsiveness to a known change, and, where a criterion claim is made, a case-mix-adjusted association that holds across the equity strata. Reliability also changes what an association means. Spearman's (1904) correction for attenuation estimates the association between true scores from the observed association and the reliabilities of the two measures, as shown in Equation 1.

$$r_{TxTy} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} \quad (1)$$

where r_{xy} is the observed association, r_{xx} and r_{yy} are the reliabilities of the indicator and the criterion, and r_{TxTy} is the estimated association between true scores.

Consider a hypothetical illustration in which two sites each observe a case-mix-adjusted correlation of 0.15 between the 48-hour initial visit indicator (I9) and 30-day unplanned hospitalization, with criterion reliability of about 0.90 for a claims-based outcome. If Site A's indicator reliability is 0.85 and Site B's is 0.45, the corrected values are about 0.17 and 0.24. The same observed figure thus carries different evidence at the two sites. Three cautions apply. Kappa is not a reliability coefficient in the classical sense, so the correction requires an intraclass or generalizability coefficient from a design that includes the relevant facets (Brennan, 2001). The correction assumes random error, whereas documentation error in a payment-linked setting may be systematic. A large correction is therefore a warning sign of bias rather than evidence of a hidden effect. The framework accordingly requires a reliability estimate beside every reported criterion association and does not permit criterion estimates to be compared across sites whose reliability differs materially.

6 PSYCHOMETRIC DECISION RULES

6.1 Content validity

Stage 1 uses a panel of at least six members, combining home health practitioners, intake and quality staff, and health-services researchers, who rate each indicator definition for relevance on a

four-point scale. Following Lynn (1986), an item-level content validity index (I-CVI) of at least 0.78 is required with six or more raters, and following Polit *et al.* (2007), the scale-level average (S-CVI/Ave) across the indicator set should reach at least 0.90. Definitions that fall below threshold are revised through a Delphi round in which consensus is defined in advance as at least 75 percent agreement, the median threshold identified in Diamond *et al.*'s (2014) review of Delphi studies. The panel also confirms that the coding rule in Table 5 can be applied from the stated data source without information outside the record. These thresholds and those that follow are summarized in Table 6. They are decision rules adopted for transparency and comparability, and a study that departs from them should state its alternative and its reason in advance.

6.2 Interrater agreement

Agreement on binary or categorical coding is reported as Cohen's kappa with its 95 percent confidence interval. Because compliance-driven items are often documented in nearly every chart, kappa can be low despite high observed agreement, the paradox described by Feinstein and Cicchetti (1990) and resolved in part by reporting additional coefficients (Cicchetti & Feinstein, 1990). Every kappa is therefore accompanied by the prevalence-adjusted bias-adjusted kappa (Byrt *et al.*, 1993) and Gwet's (2008) AC1, together with positive and negative specific agreement. Percent agreement is reported but never accepted alone. Continuous or time-based indicators use the two-way random-effects, absolute-agreement, single-rater intraclass correlation, ICC(2,1) (Koo & Li, 2016; Shrout & Fleiss, 1979). Thresholds are tiered by use. For Tier A, internal monitoring within an agency, kappa must reach 0.60 and ICC 0.75. A kappa of 0.60 marks the lower boundary of substantial agreement in the benchmarks of Landis and Koch (1977), and McHugh (2012) argues that lower values give too much room for error in health decisions; 0.60 is used as a pragmatic floor, not a natural constant. For Tier B, cross-agency comparison, public reporting, or payment-linked use, the point estimate must reach 0.80 for kappa or 0.90 for ICC, and the lower confidence bound must not fall below the Tier A floor. Where prevalence of the coded condition is below 0.10 or above 0.90, the decision rests on AC1 and specific agreement, with kappa reported for transparency.

Table 6: Decision thresholds and their sources

Property	Statistic		Tier A (internal monitoring)	Tier B (comparison, reporting, or payment-linked use)	Source
Content validity	I-CVI; CVI/Ave	S-	I-CVI at least 0.78; S-CVI/Ave at least 0.90	Same	Lynn (1986); Polit <i>et al.</i> (2007)
Definition consensus	Delphi agreement		At least 75 percent	Same	Diamond <i>et al.</i> (2014)
Categorical agreement	Kappa, PABAK, specific agreement	with AC1,	Kappa at least 0.60	Kappa at least 0.80; lower 95% bound at least 0.60	Landis and Koch (1977); McHugh (2012); Gwet (2008)

Property	Statistic	Tier A (internal monitoring)	Tier B (comparison, reporting, or payment-linked use)	Source
Continuous agreement	ICC(2,1), absolute agreement	At least 0.75	At least 0.90; lower 95% bound at least 0.75	Koo and Li (2016); Shrout and Fleiss (1979)
Agency-level reliability	Signal-to-noise ratio	Not required	At least 0.70	Adams (2009)
Differential reliability	Between-stratum difference	Reported	Difference under 0.10, or 95% CI includes zero and every stratum meets the floor	Present framework (Section 6.5)
Operational concordance	Kappa or ICC, chart versus operational record	Not required (label D)	Meets Tier B agreement threshold (label O)	Present framework (Section 6.7)

6.3 Sample size for reliability studies

Sample sizes for reliability studies should be calculated rather than assumed. For intraclass correlations, the method of Walter *et al.* (1998) gives about 49 subjects rated by two raters to distinguish an ICC of 0.80 from a null value of 0.60 with 80 percent power at a two-sided alpha of 0.05. For kappa, the goodness-of-fit approach of Donner and Eliasziw (1992) shows that required numbers rise sharply as prevalence moves away from 0.50. Table 7 reports illustrative values obtained by simulation using the large-sample variance of kappa for two raters; teams should recompute them for their own expected prevalence and kappa. The table makes one practical point plain: an equity-stratified comparison needs roughly twice as many double-reviewed charts per stratum as a single overall estimate, and more still when a compliance-driven item is almost always present. Two rural and two urban sites contributing a few dozen charts each cannot support the stratified estimates that Stage 4 requires.

Table 7: Illustrative double-review sample sizes (two raters, 80 percent power, two-sided alpha 0.05)

Objective	Expected prevalence of coded condition	Assumed values	Charts required
Lower 95% bound of kappa above 0.60	0.50	True kappa 0.80	About 80
Lower 95% bound of kappa above 0.60	0.30	True kappa 0.80	About 105

Objective	Expected prevalence of coded condition	Assumed values	Charts required
Lower 95% bound of kappa above 0.60	0.15	True kappa 0.80	About 185
Detect kappa difference between two strata	0.30	0.80 versus 0.60	About 240 per stratum
Detect kappa difference between two strata	0.15	0.80 versus 0.60	About 410 per stratum
ICC above null value (Walter <i>et al.</i> , 1998)	Not applicable	ICC 0.80 versus 0.60	About 49
Agency-level reliability of 0.70	Not applicable	Agency ICC 0.05	About 45 episodes per agency
Correlation with HHVBP TPS	Not applicable	r = 0.30	About 85 agencies

6.4 Agency-level reliability

Chart-level agreement does not establish that an agency's score distinguishes it from other agencies. When indicators are used for comparison, agency-level reliability is estimated as the share of variance in agency means attributable to true between-agency differences (Adams, 2009). With an intraclass correlation at the agency level and n episodes sampled per agency, reliability follows the Spearman-Brown relationship in Equation 2.

$$R_{agency} = \frac{n\rho}{1+(n-1)\rho} \quad (2)$$

where n is the number of episodes sampled per agency and ρ is the agency-level intraclass correlation.

Where between-agency variance is small relative to within-agency variance, so that the agency-level ICC is 0.05, about 45 episodes per agency are needed to reach 0.70 and about 76 to reach 0.80. The framework requires agency-level reliability of at least 0.70 for Tier B use, the value commonly applied in provider profiling (Adams, 2009). A generalizability study with abstractor, occasion, and agency facets yields both this quantity and the chart-level coefficients from a single design, and is the preferred approach where resources allow (Brennan, 2001). Small agencies that cannot reach the required sample within one reporting period may pool periods, provided the pooled window is stated and the indicator definition did not change within it.

6.5 Differential reliability and effect modification

The original idea of measurement invariance, tested through multi-group confirmatory factor analysis or differential item functioning, presupposes latent variables measured by several items. Most indicators here are single binary or time-based variables, so the appropriate tests are differential reliability, comparing kappa, AC1, or ICC between strata, and effect modification,

testing an indicator-by-stratum interaction in the criterion model (American Educational Research Association *et al.*, 2014). Formal invariance testing applies only to the QAPI maturity rubric (I1) if it is used as a multi-item scale (Mokkink *et al.*, 2010). Five strata are required: rurality, defined by rural-urban commuting area codes; preferred language and interpreter need (OASIS A1110); race and ethnicity (A1010, A1005); dual Medicare and Medicaid eligibility; and disability entitlement, identified from the original reason for Medicare entitlement in enrollment data. A differential is material when the between-stratum difference in agreement is at least 0.10 with a 95 percent confidence interval excluding zero, or when any stratum falls below the floor for the intended use. For effect modification, the margin is prespecified as a ratio of odds ratios between 0.80 and 1.25; an interaction whose confidence interval lies outside that range is material. Strata with too few observations are reported as unresolved, not as passing (Fashaw-Walters *et al.*, 2022).

6.6 Responsiveness

Following COSMIN terminology, the ability of an indicator to detect change in the construct it measures is termed responsiveness (Mokkink *et al.*, 2010). Where an agency implements a change to the core element of a failure mode, for example a structured referral intake checklist, the corresponding indicator should shift in the expected direction. A circularity must be avoided. If the same indicator is used both to confirm that the change was implemented and to show that the indicator responded, the two claims cannot be separated. Exposure is therefore defined from an independent record, such as an implementation log, training completion record, or system configuration date, and responsiveness is assessed against that independent exposure. An indicator that does not move after documented implementation is treated as a validity concern requiring examination of its definition before the null result is attributed to the change itself (Cronbach & Meehl, 1955).

6.7 Documentation reliability and operational reliability

Two reviewers may agree perfectly on what a chart says while the chart misstates what happened. Documentation reliability is agreement among abstractors reading the same record; operational reliability is agreement between the chart-based classification and a closer-to-real-time record of the event (Allison *et al.*, 2000). Stage 2 includes an operational verification subsample for indicators where such a record exists: electronic visit verification data for visit timing, which the 21st Century Cures Act requires for Medicaid-funded home health services (21st Century Cures Act, 2016); e-signature audit trails for order authentication (I5); scheduling timestamps for the initial visit (I9); and telephony logs for after-hours escalation (I10). The subsample is sized as in Table 7, about 105 episodes per indicator at a prevalence near 0.30, drawn at random within each participating agency. An indicator certified on documentation reliability alone is labelled D and may be used for internal monitoring and within-agency trend analysis, but not for cross-agency ranking, public reporting, or payment-linked decisions. An indicator whose chart classification meets the Tier B threshold against the operational record is labelled O and may be used for any purpose its other gates permit. Where documentation reliability is high but operational concordance is low, the indicator is measuring documentation behavior; it is relabelled as a documentation completeness indicator and its use is restricted accordingly.

6.8 Re-validation and retirement

A certification applies to the data definitions under which it was obtained. OASIS-E, effective 1 January 2023, added standardized items including ethnicity (A1005), race (A1010), language and interpreter need (A1110), health literacy (B1300), and high-risk drug classes (N0415) (CMS, 2023). OASIS-E1, effective 1 January 2025, removed several items and revised others, including

the skip pattern for the ordered start-of-care date (M0102) on which the initial visit indicator (I9) depends (CMS, 2024), and OASIS-E2 took effect on 1 April 2026. An indicator whose source item, regulatory anchor, or intake form changes is treated as provisionally unvalidated until Stage 2 evidence is refreshed, and Stage 3 evidence where the change could alter the criterion relationship. Referral completeness (I3) is triggered by changes to the agency's intake form or referral platform rather than by OASIS revisions. Every certified indicator is also reviewed at least every three years. An indicator that repeatedly fails re-validation, or whose underlying step is removed from regulation, is retired and its retirement recorded (CMS, 2026a).

7 VALIDATION PROPOSITIONS AND GUIDING PRINCIPLE

7.1 Testable propositions

Eight propositions state the empirical expectations the framework implies; each names the comparison, the threshold, and the result that would count against it. None has been tested.

Proposition 1. Within agencies that adopt structured intake and documentation standards, interrater kappa for process-reliability indicators (I3 to I6) will rise from the pre- to the post-adoption period, tested by interrupted time series or stepped-wedge design rather than by comparing adopters with non-adopters; it is supported if post-adoption kappa reaches 0.60 and the increase is at least 0.10 with a confidence interval excluding zero.

Proposition 2. Medication reconciliation accuracy (I7) will show a stronger adjusted association with medication-related emergency department visits or hospitalizations, identified from claims by reviewers blind to I7, than with all-cause 30-day hospitalization. The expectation rests on measurement proximity rather than on established effects: Redmond *et al.* (2018) found that reconciliation interventions probably make little or no difference to rehospitalization and have uncertain effects on adverse drug events, so both associations may be small. The proposition fails if the utilization association is equal or stronger.

Proposition 3. Interrater reliability of SBAR substantive completion (I8) will not differ materially between rural and urban agencies under the Section 6.5 rule; a material difference requires revision of the coding rule before cross-site use.

Proposition 4. Across at least 85 agencies, agency-level criterion-class indicators will correlate positively and moderately, between 0.30 and 0.50, with the expanded HHVBP Total Performance Score (CMS, 2021); a correlation above 0.70 would indicate redundancy with measures agencies already report.

Proposition 5. Material differential reliability for patients who need an interpreter will occur more often for indicators whose source documentation includes patient or caregiver report (I7, I8) than for indicators drawn from timestamps and practitioner records (I5, I9) (Squires *et al.*, 2022).

Proposition 6. Structural indicators will meet content validity thresholds, while their agency-level association with criterion outcomes will be smaller than that of process-class indicators; a weak association will be interpreted as the distal position of structure in Donabedian's sequence and not as evidence against content validity.

Proposition 7. Indicators tied to core elements will show responsiveness after independently documented implementation, whereas variation in adaptable elements alone will not produce consistent change.

Proposition 8. Test-retest ICC across two non-overlapping audit periods under stable practice will be higher for timestamp-derived indicators (I5, I9, I10) than for the judgment-coded SBAR indicator (I8) (Koo & Li, 2016).

7.2 Guiding principle for interpreting unsupported propositions

One further statement is a methodological principle rather than an empirical proposition, and it is stated separately for that reason. When a proposition is not supported, the indicator's operational definition and data source are examined before any conclusion is drawn that the underlying administrative process is unimportant. An unreliable measure of a real process is a measurement problem; it is not evidence that the process does not matter. The principle governs how results under Propositions 1 to 8 are read and cannot itself be falsified by data. It is consistent with Messick's (1995) view that validity concerns the inferences drawn from scores, so a failed inference prompts scrutiny of the score before scrutiny of the construct, and with Kane's (2013) requirement that each inference in an argument be evaluated in turn. Applying the principle does not license indefinite revision: an indicator that fails the same gate after two documented revisions is retired under Section 6.8.

8 STAGED VALIDATION PLAN

8.1 Stages, gates, and routes by indicator class

Table 8 sets out the four stages, their entry and exit criteria, and the propositions tested at each, and Figure 1 shows the gates and the route back to revision. Each stage gate must be passed before the next begins, so an indicator that fails early does not consume the resources later stages need. Structural indicators follow a modified route: Stage 2 for them is inter-verifier agreement, with two verifiers independently scoring the rubric or checklist from documentary evidence; Stage 3 is a descriptive association test under Proposition 6 rather than a gate; and Stage 4 checks whether verification agreement differs between rural and urban agencies. Process-reliability indicators that make no criterion claim may be certified after Stages 1, 2, and 4, with Stage 3 limited to responsiveness. Criterion indicators pass through all four stages. The sequencing reflects the principle that criterion evidence cannot substitute for reliability evidence, and that apparent criterion validity does not exempt an indicator from equity testing (Kane, 2013; Kottner *et al.*, 2011).

Table 8: Staged validation plan with gates and propositions

Stage	Primary activity	Entry criterion	Exit criterion	Propositions
1. Content review	Panel rates definitions and coding rules (Tables 4 and 5)	Draft definition complete	I-CVI at least 0.78; S-CVI/Ave at least 0.90; Delphi consensus at least 75%	None (content gate)
2. Reliability piloting	Independent double review; operational verification subsample; agency-level reliability where comparison is intended	Stage passed	1 Use-tier thresholds in Table 6 met; label D or O assigned	P1, P8

Stage	Primary activity	Entry criterion	Exit criterion	Propositions
3. Criterion, convergent, and responsiveness testing	Multilevel case-mix-adjusted HHVBP models; TPS concordance; responsiveness against independent exposure	Stage passed	2 Association and responsiveness reported with uncertainty and accompanying reliability	P2, P4, P6, P7
4. Equity-differential testing	Stratified reliability and interaction terms across five strata	Stage passed (or Stage 2 for indicators without criterion claims)	No material differential under Section 6.5, or documented revision and re-test	P3, P5

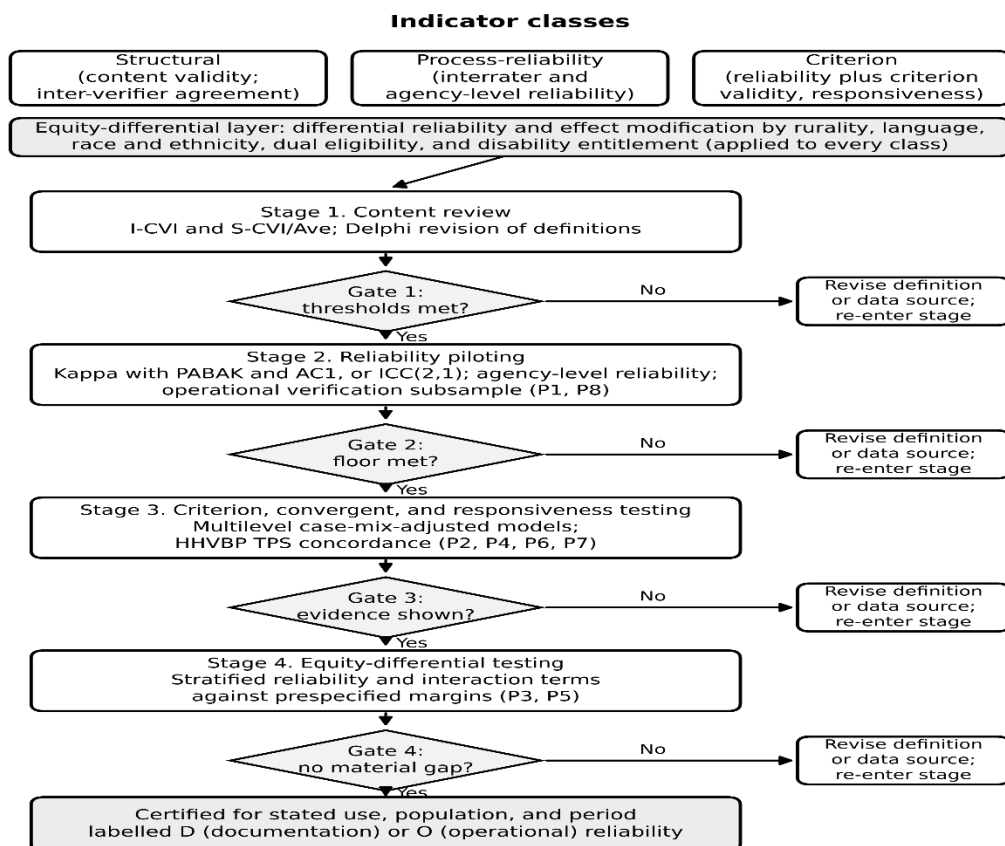


Figure 1: Indicator classes, validation stages, and decision gates. Each failed gate returns the indicator to definition or data-source revision and re-entry at the same stage; certification is specific to the stated use, population, and period, and carries a D (documentation) or O (operational) reliability label.

8.2 Sites and samples for stratified estimates

Stratified estimates require planned sampling. A collaborative intending Tier B use should sample double-reviewed charts within each stratum at the numbers in Table 7, oversampling strata that are small in the overall caseload, such as patients who need an interpreter, rather than relying on their natural frequency. Because agency documentation practice is itself a source of variance, the charts in each stratum should come from several agencies; as a practical minimum, the framework recommends that each rural and urban stratum include at least ten agencies, so that between-agency variance within the stratum can be estimated (Brennan, 2001). Sample sizes for kappa should be calculated with the method of Donner and Eliasziw (1992) and for intraclass correlations with that of Walter *et al.* (1998), using the expected prevalence and agreement for each indicator. A collaborative of ten agencies can complete Stages 1 and 2 and the rurality comparison but will usually lack the numbers for language, race and ethnicity, and disability strata; those comparisons are then reported as unresolved.

8.3 Criterion design

Episodes are nested within agencies, so criterion associations are estimated with multilevel models that include random agency intercepts, and agency-level characteristics such as size and rurality are entered as fixed effects. Case-mix adjustment uses the risk factors CMS specifies for home health outcome measures, including age, payer, admission source, the risk-for-hospitalization items in M1033, living situation, cognitive and functional status, and hierarchical condition categories (CMS, 2026e). Where the criterion is a claims-based utilization outcome, the adjustment approach follows the corresponding CMS claims-based specifications (Abt Associates & Acumen, 2023; Acumen, 2012). The HHVBP criterion is the Total Performance Score of the expanded model, which began with performance year 2023 (CMS, 2021). Because that score is annual and agency-level, Proposition 4 is tested at the agency level and needs about 85 agencies to detect a correlation of 0.30 with 80 percent power. For Proposition 2, contamination is avoided by ascertaining medication-related events from claims diagnosis codes coded by reviewers who do not see the agency's reconciliation record, so the criterion does not depend on the same discrepancy identification that produces the indicator.

8.4 A hypothetical worked example

The following example is hypothetical and illustrates the procedure only. A ten-agency collaborative applies the framework to face-to-face encounter documentation (I6) for traditional Medicare fee-for-service episodes. In Stage 1, a panel of eight rates the definition; the I-CVI for the coding rule falls to 0.75 until the rule is revised to state which practitioners may perform the encounter and that permitted telehealth encounters qualify, as 42 CFR 424.22(a)(1)(v) provides (CMS, 2026b). In Stage 2, reliability meets the Tier A floor overall, but one agency shows low agreement, concentrated in encounters documented in a hospital record that the agency held only as a summary. The coding rule is revised to specify acceptable source documents and the agency is re-piloted. One discordant agency cannot show that the problem is specific to rural agencies or to any payer; that question belongs to Stage 4. Stage 3 finds a modest adjusted association between I6 and on-time plan-of-care initiation. Stage 4 finds no material rural-urban differential, but the language and disability strata are too small for a conclusion. The indicator is certified for Tier A use with label D, and the unresolved strata are reported.

8.5 Reporting partial validation

Validation will proceed unevenly across indicators and sites, and partial validation should be reported as partial. Each indicator's status is reported as the highest stage passed, the use tier, the D or O label, the strata tested and those unresolved, the data-definition version, and the date. An indicator that has passed Stages 1 and 2 but lacks linked outcome data is reported as reliability-certified with criterion validity untested, not rounded up to validated. Reliability studies should follow the Guidelines for Reporting Reliability and Agreement Studies (Kottner *et al.*, 2011), and measurement properties should be named and reported using COSMIN terminology (Mokkink *et al.*, 2010), so that results from different collaboratives can be compared. Smaller agencies without access to linked outcome data can still complete Stages 1 and 2 and contribute to multi-agency or association-led efforts for Stages 3 and 4, which are the more plausible venue for criterion and equity testing (Adams, 2009).

9 DISCUSSION

9.1 What validation establishes and what it does not

A certified indicator, in this framework, has documented reliability at the level of its intended use, content validity, and, for criterion indicators, a case-mix-adjusted association with an independently ascertained outcome that holds across the tested strata. This supports one kind of inference: that a difference observed between two agencies, or between two periods at one agency, probably reflects a difference in the administrative process rather than abstraction noise or unequal documentation. It does not show that improving the indicator causes better outcomes, which requires an intervention design, and it does not confer regulatory endorsement. Kane's (2013) framing makes the limit precise: the validation argument supports the stated interpretation and use and nothing beyond them. Messick's (1995) consequential dimension adds that an indicator used beyond its certified tier can cause harm even if its measurement properties are sound, for example by ranking agencies on an indicator certified only for internal monitoring.

9.2 Compliance incentives and ceiling effects

Several indicators map onto requirements with payment consequences, so agencies have strong incentives to document them completely. That incentive produces two measurement risks the framework addresses directly. First, near-universal documentation pushes prevalence toward ceiling and makes kappa unstable, which is why Section 6.2 requires prevalence-robust coefficients and specific agreement (Byrt *et al.*, 1993; Gwet, 2008). Second, consistent documentation can coexist with inconsistent practice, which is why Section 6.7 separates documentation from operational reliability and restricts the uses of indicators that have not been verified against operational records. Neither risk is resolved by larger samples alone. The operational verification subsample is the main safeguard, and the D and O labels make the extent of that verification visible to anyone using the indicator (Allison *et al.*, 2000).

9.3 Why the framework does not propose a composite score

A single composite administrative reliability score would be convenient for governance reporting, but the framework does not propose one. The NQF composite guidance requires that each component and the composite itself be shown reliable and valid, and that the weighting and aggregation method be justified (National Quality Forum, 2013). The indicators here are formative: each describes a distinct step that contributes to administrative reliability, and they need not correlate with one another. Internal consistency coefficients such as Cronbach's alpha therefore do not indicate the quality of such a set, and the framework does not use them except for the QAPI

maturity rubric when it is scored as a scale (Mokkink *et al.*, 2010). A collaborative that wants a summary view should build it only from components certified for the same use tier, report each component's status alongside it, and validate the composite as a measure in its own right.

9.4 Generalizability beyond the Medicare context

Several sources used for mechanism-level support originate outside the United States Medicare home health system, including an Italian discharge-planning comparison (Damiani *et al.*, 2009), an Ethiopian medication-reconciliation study (Tamiru *et al.*, 2018), and an Argentine readmission cohort among older adults discharged with home care (Cunha Ferré *et al.*, 2019). They support the plausibility that communication, reconciliation, and coordination processes relate to readmission-type outcomes in many systems. The operational definitions in Tables 4 and 5, however, are anchored to OASIS, the Conditions of Participation, and HHVBP. A team adapting the framework elsewhere should use Stage 1 to re-derive definitions from its own regulatory and documentation equivalents, and should treat the thresholds in Table 6 as starting points subject to the same justification.

9.5 Limitations and future work

The framework has not been tested, and its thresholds are decision rules whose consequences will become clear only through use; a study that finds them too strict or too lenient for a particular indicator should report that finding. The literature synthesis was structured but not systematic, so relevant measurement studies may have been missed (Vassar & Holzmann, 2013). Operational verification depends on records, such as electronic visit verification and telephony logs, that not every agency holds for every payer. The equity strata depend on the completeness of OASIS demographic items and enrollment data, and small strata will often remain unresolved in single collaboratives. Three extensions follow. The first is an empirical Stage 1 and Stage 2 study of the ten indicators across agencies varying in size and rurality. The second is a reduced version of the framework for single-agency QAPI use. The third is application of the same class and gate logic to clinical process measures such as care-plan appropriateness (Fashaw-Walters *et al.*, 2022; Kottner *et al.*, 2011).

10 Conclusion

Administrative reliability indicators are already used to audit staff, direct improvement effort, and compare home health agencies, often without evidence that they measure consistently or fairly. This paper sets out what that evidence should be. It defines ten indicators operationally, assigns each to a class with a stated validity requirement, and applies an equity-differential analysis to all of them. It sets numerical decision rules for content validity, interrater and agency-level reliability, and differential reliability, tiered by how the indicator will be used. It separates reliability of documentation from reliability of practice and limits the uses of indicators verified only on paper. Eight propositions and a four-stage pathway make the framework testable, and a separately stated principle governs how failed propositions are read. The framework is intended to be applied, revised, and, where the evidence requires, corrected by the empirical studies it is designed to guide.

REFERENCES

1. Abt Associates, & Acumen. (2023). Specifications for the home health within-stay potentially preventable hospitalization measure. Centers for Medicare & Medicaid Services. <https://www.cms.gov/files/document/hhqrpspecificationspotentiallypreventablehospitalizations.pdf>
2. Acumen. (2012). Specifications for home health claims-based utilization measures. Centers for Medicare & Medicaid Services. <https://edit.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Downloads/ClaimsBasedUtilizationMeasuresSpecifications.pdf>

3. Adams, J. L. (2009). The reliability of provider profiling: A tutorial (Technical Report TR-653-NCQA). RAND Corporation. https://www.rand.org/pubs/technical_reports/TR653.html
4. Allison, J. J., Wall, T. C., Spettell, C. M., Calhoun, J., Fargason, C. A., Kobylinski, R., Farmer, R., & Kiefe, C. I. (2000). The art and science of chart review. *The Joint Commission Journal on Quality Improvement*, 26(3), 115-136. [https://doi.org/10.1016/S1070-3241\(00\)26009-4](https://doi.org/10.1016/S1070-3241(00)26009-4)
5. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.
6. Anderson, M. A., Clarke, M. M., Helms, L. B., & Foreman, M. D. (2005). Hospital readmission from home health care before and after prospective payment. *Journal of Nursing Scholarship*, 37(1), 73-79. <https://doi.org/10.1111/j.1547-5069.2005.00001.x>
7. Bartzak, P. J. (2022). Implications of incomplete medication reconciliation during transitions of care. *MEDSURG Nursing*, 31(2), 133. <https://doi.org/10.62116/msj.2022.31.2.133>
8. Battelle. (2024). Endorsement and maintenance (E&M) guidebook (Version 2.1). Partnership for Quality Measurement. https://p4qm.org/sites/default/files/2024-08/Del-3-6-Endorsement-and-Maintenance-Guidebook-Final_0.pdf
9. Becerra, A. (2019). Impact of medication reconciliation programs on reducing readmission rate and enhancing medication safety during transitions of care. *The Permanente Journal*. <https://doi.org/10.7812/tpp/19-039-27>
10. Boling, P. A., & Parsons, P. (2007). A research and policy agenda for transitions from nursing homes to home. *Home Health Care Services Quarterly*, 26(4), 121-131. https://doi.org/10.1300/j027v26n04_09
11. Brennan, R. L. (2001). Generalizability theory. Springer. <https://doi.org/10.1007/978-1-4757-3456-0>
12. Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429. [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
13. Centers for Medicare & Medicaid Services. (2021). Medicare and Medicaid programs; CY 2022 home health prospective payment system rate update; home health value-based purchasing model requirements and model expansion. *Federal Register*, 86(214), 62240-62431. <https://www.federalregister.gov/d/2021-23993>
14. Centers for Medicare & Medicaid Services. (2023). Outcome and Assessment Information Set OASIS-E guidance manual (effective January 1, 2023). Retrieved October 9, 2026, from <https://www.cms.gov/medicare/quality/home-health/oasis-user-manuals>
15. Centers for Medicare & Medicaid Services. (2024). Outcome and Assessment Information Set OASIS-E1 guidance manual (effective January 1, 2025). Retrieved October 9, 2026, from <https://www.cms.gov/medicare/quality/home-health/oasis-user-manuals>
16. Centers for Medicare & Medicaid Services. (2026a). Blueprint measure lifecycle. CMS Measures Management System Hub. Retrieved October 9, 2026, from <https://mmshub.cms.gov/blueprint-measure-lifecycle-overview>
17. Centers for Medicare & Medicaid Services. (2026b). Certification and plan requirements, 42 C.F.R. § 424.22. *Electronic Code of Federal Regulations*. Retrieved October 9, 2026, from <https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-B/part-424/subpart-B/section-424.22>
18. Centers for Medicare & Medicaid Services. (2026c). Home health services, 42 C.F.R. pt. 484. *Electronic Code of Federal Regulations*. Retrieved October 9, 2026, from <https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-484>
19. Centers for Medicare & Medicaid Services. (2026d). OASIS user manuals. Retrieved October 9, 2026, from <https://www.cms.gov/medicare/quality/home-health/oasis-user-manuals>
20. Centers for Medicare & Medicaid Services. (2026e). Risk adjustment technical steps and risk factor specifications (effective January 1, 2025). Retrieved October 9, 2026, from <https://www.cms.gov/files/document/risk-adjustment-technical-specifications-2025.pdf-0>
21. Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551-558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
22. Clark, J. (1991). HIV nursing management in the home health care setting. *Journal of Home Health Care Practice*, 3(2), 1-16. <https://doi.org/10.1177/108482239100300203>
23. Coleman, E. A., Parry, C., Chalmers, S., & Min, S.-J. (2006). The care transitions intervention: Results of a randomized controlled trial. *Archives of Internal Medicine*, 166(17), 1822-1828. <https://doi.org/10.1001/archinte.166.17.1822>
24. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/h0040957>

25. Cunha Ferré, M. F., Gallo Acosta, C. M., Dawidowski, A. R., Senillosa, M. B., Scozzafava, S. M., & Saimovici, J. M. (2019). 72-hour hospital readmission of older people after hospital discharge with home care services. *Home Health Care Services Quarterly*, 38(3), 153-161. <https://doi.org/10.1080/01621424.2019.1616024>
26. Damiani, G., Federico, B., Venditti, A., Sicuro, L., Rinaldi, S., Cirio, F., Pregno, C., & Ricciardi, W. (2009). Hospital discharge planning and continuity of care for aged people in an Italian local health unit: Does the care-home model reduce hospital readmission and mortality rates? *BMC Health Services Research*, 9, 22. <https://doi.org/10.1186/1472-6963-9-22>
27. Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67(4), 401-409. <https://doi.org/10.1016/j.jclinepi.2013.12.002>
28. Distelhorst, K. S., & Hansen, D. M. (2022). Neighborhood matters for transitional care and heart failure hospital readmission in older adults. *Geriatric Nursing*, 47, 183-190. <https://doi.org/10.1016/j.gerinurse.2022.07.014>
29. Donabedian, A. (1988). The quality of care: How can it be assessed? *JAMA*, 260(12), 1743-1748. <https://doi.org/10.1001/jama.1988.03410120089033>
30. Donner, A., & Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine*, 11(11), 1511-1519. <https://doi.org/10.1002/sim.4780111109>
31. Fashaw-Walters, S. A., Rahman, M., Gee, G., Mor, V., White, M., & Thomas, K. S. (2022). Out of reach: Inequities in the use of high-quality home health agencies. *Health Affairs*, 41(2), 247-255. <https://doi.org/10.1377/hlthaff.2021.01408>
32. Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543-549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
33. Gerace, L. M. (1990). Nursing assessment in psychiatric home health care. *Journal of Home Health Care Practice*, 2(3), 9-15. <https://doi.org/10.1177/108482239000200304>
34. Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29-48. <https://doi.org/10.1348/000711006X126600>
35. Haggerty, J. L., Reid, R. J., Freeman, G. K., Starfield, B. H., Adair, C. E., & McKendry, R. (2003). Continuity of care: A multidisciplinary review. *BMJ*, 327(7425), 1219-1221. <https://doi.org/10.1136/bmj.327.7425.1219>
36. Hansen, L. O., Williams, M. V., & Singer, S. J. (2011). Perceptions of hospital safety climate and incidence of readmission. *Health Services Research*, 46(2), 596-616. <https://doi.org/10.1111/j.1475-6773.2010.01204.x>
37. Heeke, S., Wood, F., & Schuck, J. (2014). Improving care transitions from hospital to home. *Journal of Nursing Care Quality*, 29(2), E21-E28. <https://doi.org/10.1097/ncq.0b013e3182a520b6>
38. Heffner, C., Dillaman, M., & Hill, J. (2020). Pharmacist-driven medication reconciliation reduces oral oncolytic medication errors during transitions of care. *American Journal of Health-System Pharmacy*, 77(Supplement 4), S100-S104. <https://doi.org/10.1093/ajhp/zxaa168>
39. Hung, D., & Leidig, R. C. (2015). Implementing a transitional care program to reduce hospital readmissions among older adults. *Journal of Nursing Care Quality*, 30(2), 121-129. <https://doi.org/10.1097/ncq.0000000000000091>
40. Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press. <https://doi.org/10.17226/10027>
41. Jack, B. W., Chetty, V. K., Anthony, D., Greenwald, J. L., Sanchez, G. M., Johnson, A. E., Forsythe, S. R., O'Donnell, J. K., Paasche-Orlow, M. K., Manasseh, C., Martin, S., & Culpepper, L. (2009). A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. *Annals of Internal Medicine*, 150(3), 178-187. <https://doi.org/10.7326/0003-4819-150-3-200902030-00007>
42. Joo, J. Y., & Liu, M. F. (2023). Transitional care interventions for supporting frail older adults discharged from hospitals: An umbrella review. *Geriatric Nursing*, 50, 80-89. <https://doi.org/10.1016/j.gerinurse.2022.12.021>
43. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
44. Kellogg, F. R., Brickner, P. W., Conley, L., & Conroy, M. (1991). Controlling hospital readmission of elderly persons living at home. *Home Health Care Services Quarterly*, 12(2), 5-16. https://doi.org/10.1300/j027v12n02_02

45. Kinatukara, S., Rosati, R. J., & Huang, L. (2005). Assessment of OASIS reliability and validity using several methodological approaches. *Home Health Care Services Quarterly*, 24(3), 23-38. https://doi.org/10.1300/J027v24n03_02
46. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>
47. Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96-106. <https://doi.org/10.1016/j.jclinepi.2010.03.002>
48. Kripalani, S., LeFevre, F., Phillips, C. O., Williams, M. V., Basaviah, P., & Baker, D. W. (2007). Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care. *JAMA*, 297(8), 831-841. <https://doi.org/10.1001/jama.297.8.831>
49. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>
50. Li, H., Morrow-Howell, N., & Proctor, E. K. (2004). Post-acute home care and hospital readmission of elderly patients with congestive heart failure. *Health & Social Work*, 29(4), 275-285. <https://doi.org/10.1093/hsw/29.4.275>
51. Lodge, M., Bashiri, S., Durazo-Arvizu, R., & Russell, C. J. (2024). The association between home health nursing and 30-day hospital readmission after tracheostomy in children. *Home Healthcare Now*, 42(6), 323-330. <https://doi.org/10.1097/nhh.0000000000001294>
52. Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-385. <https://doi.org/10.1097/00006199-198611000-00017>
53. Madigan, E. A., & Fortinsky, R. H. (2004). Interrater reliability of the Outcomes and Assessment Information Set: Results from the field. *The Gerontologist*, 44(5), 689-692. <https://doi.org/10.1093/geront/44.5.689>
54. Magilvy, J. K., & Lakomy, J. M. (1992). Transitions of older adults to home care. *Home Health Care Services Quarterly*, 12(4), 59-70. https://doi.org/10.1300/J027v12n04_05
55. Martens, K. H. (2000). Home care nursing for persons with congestive heart failure: Description and relationship to hospital readmission. *Home Healthcare Nurse*, 18(6), 404-409. <https://doi.org/10.1097/00004045-200006000-00018>
56. Martin, C. M. (2012). Avoiding errors during transitions of care: Medication reconciliation. *The Consultant Pharmacist*, 27(11), 764-769. <https://doi.org/10.4140/tcp.n.2012.764>
57. McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276-282. <https://doi.org/10.11613/BM.2012.031>
58. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066x.50.9.741>
59. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737-745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
60. Narayan, M. C., & Scafide, K. N. (2017). Systematic review of racial/ethnic outcome disparities in home health care. *Journal of Transcultural Nursing*, 28(6), 598-607. <https://doi.org/10.1177/1043659617700710>
61. National Quality Forum. (2013). Composite performance measure evaluation guidance. https://www.qualityforum.org/Publications/2013/04/Composite_Performance_Measure_Evaluation_Guidance.aspx
62. National Quality Forum. (2021). Measure evaluation criteria and guidance for evaluating measures for endorsement. https://www.qualityforum.org/Measuring_Performance/Submitting_Standards/2021_Measure_Evaluation_Criteria_and_Guidance.aspx
63. Navarro, A. E., Enguidanos, S., & Wilber, K. H. (2012). Identifying risk of hospital readmission among Medicare aged patients: An approach using routinely collected data. *Home Health Care Services Quarterly*, 31(2), 181-195. <https://doi.org/10.1080/01621424.2012.681561>
64. Naylor, M. D., Aiken, L. H., Kurtzman, E. T., Olds, D. M., & Hirschman, K. B. (2011). The care span: The importance of transitional care in achieving health reform. *Health Affairs*, 30(4), 746-754. <https://doi.org/10.1377/hlthaff.2011.0041>

65. Naylor, M. D., Brooten, D. A., Campbell, R. L., Maislin, G., McCauley, K. M., & Schwartz, J. S. (2004). Transitional care of older adults hospitalized with heart failure: A randomized, controlled trial. *Journal of the American Geriatrics Society*, 52(5), 675-684. <https://doi.org/10.1111/j.1532-5415.2004.52202.x>
66. Nuckols, T. K. (2015). County-level variation in readmission rates: Implications for the Hospital Readmission Reduction Program's potential to succeed. *Health Services Research*, 50(1), 12-19. <https://doi.org/10.1111/1475-6773.12268>
67. Ogundipe, M., & Ominde, S. (2024). Administrative governance as a missing link in continuity of care: Policy recommendations for modernizing home health services. Manuscript submitted for publication.
68. Ogundipe, M., & Ominde, S. (2025). Improving continuity of care through standardized care transition protocols in Medicare-certified home health agencies. Manuscript submitted for publication.
69. Pesko, M. F., Gerber, L. M., Peng, T. R., & Press, M. J. (2018). Home health care: Nurse-physician communication, patient severity, and hospital readmission. *Health Services Research*, 53(2), 1008-1024. <https://doi.org/10.1111/1475-6773.12667>
70. Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459-467. <https://doi.org/10.1002/nur.20199>
71. Proctor, E. K., Morrow-Howell, N., Li, H., & Dore, P. (2000). Adequacy of home care and hospital readmission for elderly congestive heart failure patients. *Health & Social Work*, 25(2), 87-96. <https://doi.org/10.1093/hsw/25.2.87>
72. Redmond, P., Grimes, T. C., McDonnell, R., Boland, F., Hughes, C., & Fahey, T. (2018). Impact of medication reconciliation for improving transitions of care. *Cochrane Database of Systematic Reviews*, 2018(8), CD010791. <https://doi.org/10.1002/14651858.CD010791.pub2>
73. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. <https://doi.org/10.1037/0033-2909.86.2.420>
74. Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72-101. <https://doi.org/10.2307/1412159>
75. Squires, A., Ma, C., Miner, S., Feldman, P., Jacobs, E. A., & Jones, S. A. (2022). Assessing the influence of patient language preference on 30 day hospital readmission risk from home health care: A retrospective analysis. *International Journal of Nursing Studies*, 125, 104093. <https://doi.org/10.1016/j.ijnurstu.2021.104093>
76. Tamiru, A., Edessa, D., Sisay, M., & Mengistu, G. (2018). Magnitude and factors associated with medication discrepancies identified through medication reconciliation at care transitions of a tertiary hospital in eastern Ethiopia. *BMC Research Notes*, 11(1). <https://doi.org/10.1186/s13104-018-3668-z>
77. Thobaben, M. (2012). Psychiatric home health care skilled nursing services. *Home Health Care Management & Practice*, 25(1), 32-34. <https://doi.org/10.1177/1084822312461354>
78. Treece, J., Ghouse, M., Rashid, S., Arikapudi, S., Sankhyani, P., Kohli, V., O'Neill, L., Addo-Yobo, E., Bhattad, V., & Baumrucker, S. J. (2018). The effect of hospice on hospital admission and readmission rates: A review. *Home Health Care Management & Practice*, 30(3), 140-146. <https://doi.org/10.1177/1084822318761105>
79. 21st Century Cures Act, Pub. L. No. 114-255, § 12006, 130 Stat. 1033 (2016).
80. Vassar, M., & Holzmann, M. (2013). The retrospective chart review: Important methodological considerations. *Journal of Educational Evaluation for Health Professions*, 10, 12. <https://doi.org/10.3352/jeehp.2013.10.12>
81. Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1), 101-110. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980115\)17:1<101::AID-SIM727>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E)