

Prove the Action Happened Acceptance Tests for Healthcare Agent Suppliers

A Manager Method for Verifying Completed Work

Phil Hatch | Akholi

October 23, 2026

Healthcare delivery managers should accept agent work through the authoritative result in the receiving workflow. Fluent text, tool-call success, and benchmark performance cannot by themselves prove that the purchased administrative action was completed.

Define Completion Where the Work Must Arrive

The healthcare service manager should define each purchased result in the system or process that relies on it. A prepared referral, a transmitted referral, and a referral accepted by the intended recipient are different states. The supplier's completion rule should identify which state the contract requires and what evidence confirms it.

Stanford's medical-AI evaluation distinguishes task execution from more limited benchmark performance in simulated settings.¹ Managers should use that distinction to design local acceptance, without treating research benchmark results as clinical deployment rates or a prediction of this supplier's performance.

¹ Nigam Shah and Jonathan Chen, "Measuring What Medical AI Can Actually Do," Stanford University, October 19, 2026, "PhysicianBench and MedAgentBench" and "HealthAdminBench," <https://hai.stanford.edu/news/measuring-what-medical-ai-can-actually-do>.

Begin with one workflow and identify the authoritative evidence for each consequential action. A screenshot of a draft or a model narrative may show preparation. It cannot establish that the intended case, document, or message reached the correct destination and produced the required state.

Build Tests Around the Whole Case

The study describes reported completion without the corresponding result and manual work used to make delivery succeed.² Those banking observations suggest a test mechanism, not a healthcare failure rate. The buyer should observe the actual receiving workflow and preserve human interventions in the result.

For a hypothetical appointment-support service, the test should distinguish a slot search, a booking request, a confirmed appointment, and a patient notification. Include a changed schedule, unavailable slot, incorrect patient identifier, and failed acknowledgment. A service that succeeds only on the normal path has not demonstrated its exception handling.

Use representative test cases under appropriate privacy and clinical controls. Record the intended result before the run so the supplier cannot redefine success after seeing the output. Where synthetic cases simplify the environment, state which production conditions remain untested.

Use a Joint Acceptance Evidence Record

Figure 1 proposes a record linking requested action, observed external result, exception disposition, and final acceptance. Give each case a stable reference across buyer and supplier systems. Preserve the relevant configuration and policy version so the result can be interpreted after a release changes the service.

A human-assisted success should be recorded as such. The case may be acceptable operationally while failing a claim of unattended completion. Record the reason for intervention, the person's authority, work performed, and final result. This distinction supports both reliable service and honest economics.

Test adverse conditions deliberately: missing information, duplicate requests, unavailable tools, ambiguous instructions, and recovery after interruption. Confirm that the service does not create an invented completion or take an unapproved clinical action to close the case. Unresolved cases should remain visible with an owner and next step.

² Phil Hatch, *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*, version 1.8.2 (Akholi, forthcoming manuscript dated November 1, 2026), Part 3, AI Technology, sec. 4.1, "Accuracy of the Tools' Own Success and Failure Reports," indicator C172, <https://doi.org/10.2139/ssrn.7164559>; Hatch, *Agentic AI Outsourcing*, Part 5, "The Five Manual Alteration Methods" and "Disclosure of Manual Alteration," table 5.8.

Figure 1. The healthcare action acceptance record

Requested action	External result	Human assistance	Exception
Decision	Decision	Decision	Decision
What must happen and where	Did the receiving workflow change	Who finished or corrected the work	What remains unresolved
Evidence	Evidence	Evidence	Evidence
Predefined case and result	Authoritative record and acknowledgment	Intervention and authority record	Owner next action and final disposition
Maintain one reconciled operating record across buyer and provider. Investigate gaps, assign corrective authority, and repeat the affected acceptance test.			

Author's proposed method; adapt to the service and applicable jurisdiction.

Source: Phil Hatch, proposed framework developed for this paper.

Preserve Clinical and Local Safety Boundaries

NHS England's review describes distinct clinical-risk responsibilities for manufacturers and deploying organizations under DCB0129 and DCB0160.³ That consultation context is not a global standard. A healthcare buyer should determine the current requirements for its own function, institution, and jurisdiction.

EU, UK, Swiss, and US services may require different clinical, privacy, device, and professional-duty assessments. An administrative contract should identify which clinical decisions it does not authorize and how uncertain cases reach a qualified person. A successful local workflow test does not remove those obligations.

The manager should connect acceptance evidence to ongoing monitoring. A new model, tool, recipient interface, or patient population can invalidate part of the earlier test. Identify the affected cases and require proportionate reassessment rather than assume the original demonstration covers every later configuration.

³ NHS England, "National Review of Clinical Risk Management Standards DCB0129 and DCB0160: Supporting Information," NHS England, June 29, 2026, "Why the standards were introduced"; consultation context, <https://www.england.nhs.uk/long-read/national-review-of-clinical-risk-management-standardsdcbo129-and-dcbo160-supporting-information/>.

Approve the Result the Buyer Can Verify

A reasonable objection is that observing every action may be costly. Use risk-based monitoring and sampling where appropriate, while retaining case-level evidence for material actions and investigation. Sampling is a coverage decision that should be stated clearly; it should not convert unobserved work into verified completion.

Before production acceptance, reconcile normal, failed, corrected, and interrupted cases with the receiving systems. Resolve discrepancies, assign corrective owners, and repeat the affected tests. Report accepted outcomes, human assistance, unresolved work, and test limitations separately.

The manager should accept the service when the purchased action can be shown to have occurred under the agreed conditions. The evidence should survive supplier boundaries and later questions. An output that sounds complete is useful only when the healthcare workflow confirms the result on which people will rely.

Evidence and Source Note

The study is a nonrepresentative cross-section of 850 vendor delivery-team respondents on 150 engagements serving Tier 1 financial institutions, collected during January–June 2026. Neither banks nor vendors corroborated the accounts. It establishes reported experience, not causation, legal breach, or market prevalence. Its application here is a mechanism to investigate, not an estimate of conditions in other industries. The author-supplied version 1.8.2 is dated November 1, 2026 and is cited as a forthcoming manuscript consulted October 23, 2026.⁴

This is an analytical decision paper based on the supplied study and selected primary public sources checked through October 23, 2026. The original contribution is a healthcare action acceptance record that separates generated content, externally verified execution, human-assisted completion, and unresolved case state. The proposed method has not been independently validated. Legal requirements depend on the entity, activity, jurisdiction, and effective provisions; supervisory observations, research, and professional commentary have distinct authority.

Bibliography

Hatch, Phil. *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*. Version 1.8.2. Akholi. Forthcoming manuscript dated November 1, 2026. Author-supplied manuscript consulted October 23, 2026. <https://doi.org/10.2139/ssrn.7164559>.

NHS England. “National Review of Clinical Risk Management Standards DCB0129 and DCB0160: Supporting Information.” NHS England. June 29, 2026. <https://www.england.nhs.uk/long-read/national-review-of-clinical-risk-management-standardsdcb0129-and-dcb0160-supporting-information/>.

Shah, Nigam, and Jonathan Chen. “Measuring What Medical AI Can Actually Do.” Stanford University. October 19, 2026. <https://hai.stanford.edu/news/measuring-what-medical-ai-can-actually-do>.

⁴ Hatch, *Agentic AI Outsourcing*, Introduction, “Executive Summary” and “Evidentiary Basis and Its Limits.”