Clinical AI Toolkit

May 2026

CHIEF | Executive Forum

Digital Health CANADA

Contents

Introduction

Despite rapid AI adoption across Canadian health systems as documented in our environmental scan, there is no standardized, proportionate, lifecycle-based evaluation designed for delivery leaders. Existing frameworks focus on technical validation or governance intake, leaving a gap in practical scale decision support. As a result, many AI initiatives stall between pilot and scale due to unclear evidence thresholds, fragmented governance, and misaligned financial accountability.

How to use this toolkit

The CHIEF Executive Forum AI Toolkit ("the toolkit") supports consistent, practical, and scalable assessment of clinical AI implementations across Canadian healthcare settings.

The toolkit brings together key elements needed to evaluate early impact, interpret results, and support responsible scaling. It provides a minimum evaluation that any organization can use, including those with limited analytic capacity, along with optional expanded components for teams that wish to conduct deeper analysis.

The toolkit supports structured decision-making at four key inflection points:

- Pre-implementation readiness assessment
- Early-stage evaluation* and directional signal detection
- Scale decision and investment confirmation
- Ongoing monitoring and sustainment

The toolkit is designed to support practical, low-burden evaluation of clinical AI implementations. It provides a structured approach to generating early signals of impact and supporting decisions about whether to continue, refine, or scale an AI tool.

A typical approach is:

- Start with the Minimum Viable Evaluation (MVE) to capture a simple before-and-after assessment using a small number of indicators.
- Select 2 to 4 relevant metrics from the KPI Library to structure measurement across key domains.
- Use the ROI approach to interpret early value signals, including time savings, avoided events, and workflow impact.
- Refer to Case Examples to guide interpretation and understand how similar evaluations may be applied in practice.
- Expand the evaluation only if needed, using additional KPIs or more detailed analysis for higher-risk or higher-stakes implementations.

Organizations should focus on a small number of meaningful indicators rather than attempting comprehensive measurement. The goal is to generate clear, directional insight to support decision-making, not to produce definitive proof of impact.

Who should use this toolkit?

The toolkit is intended for Chief Information Officers, Chief Medical Information Officers, clinical program leaders, digital innovation executives, and quality leaders responsible for evaluating AI-enabled workflows within healthcare delivery organizations.

*For the purposes of this toolkit, early-stage evaluation refers to pilot or early rollout periods in which organizations are assessing safety, usability, workflow impact, and early value signals, rather than attempting full outcome validation or mature long-term ROI analysis.

Guardrails + Intended Use

This toolkit is designed to support consistent, practical evaluation of clinical AI implementations. It provides structure and shared language but does not prescribe a single methodology or replace formal health technology assessment processes.

Security, privacy, and regulatory compliance are assumed to be addressed through existing organizational governance and procurement processes and are not evaluated within the scope of this toolkit.

The toolkit also does not require technical model validation or algorithm auditing. Oversight checks focus on safe and appropriate use within the clinical workflow.

Evaluation Is Not Validation

Completion of the evaluation does not constitute regulatory approval, certification, or validation of safety or effectiveness. Organizations remain responsible for meeting applicable regulatory, privacy, and clinical governance requirements.

Directional ROI

Financial and operational estimates generated through this toolkit are intended to be directional and context dependent. Results should be interpreted relative to the defined perspective, comparator, and time horizon.

Minimum Means Minimum

The Minimum Viable Evaluation (MVE) is intentionally pragmatic. It supports early learning and structured reflection in low-capacity settings. Organizations with greater analytic capacity may conduct deeper analyses but the absence of advanced methods does not invalidate early evaluation.

Ongoing Oversight

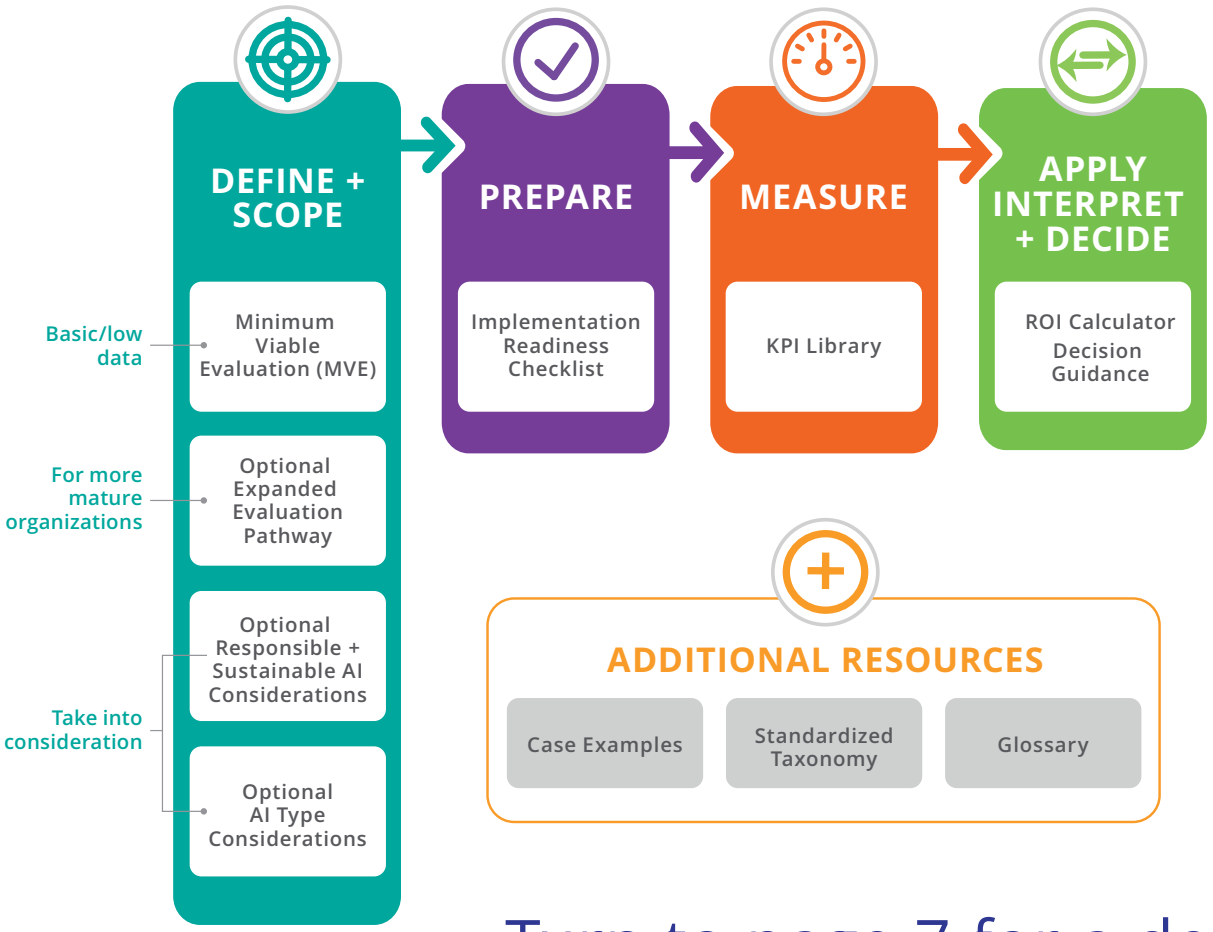
AI implementation is not a one-time event. Performance, workflow impact, and safety signals should be revisited periodically as contexts evolve.

Clinical AI Evaluation Process: Components

The toolkit components are designed to work together:

- The Minimum Viable Evaluation generates early signals of impact
- The KPI Library helps structure what is measured
- The ROI approach interprets these signals in terms of value

Together, these elements support decisions about whether to continue, refine, or scale an AI tool.



Turn to page 7 for a deeper dive into these components.

Clinical AI Evaluation Process: Component Details

The sections below align with the phases of the Clinical AI Evaluation Process on the previous page. Colour coding reflects this alignment.

DEFINE and SCOPE

The **Minimum Viable Evaluation** provides a structured before–after assessment, a brief clinician experience check, and early directional ROI signals. The goal is to generate shared understanding, not to produce definitive evidence.

Organizations with greater analytic capacity or higher-risk implementations may use the **Expanded Evaluation** and **Simple Evaluation** to deepen analysis, conduct scenario modeling, or perform trend review.

PREPARE

Use the **Implementation Readiness Checklist** to ensure workflow fit, data basics, and staff understanding are in place. This helps interpret early results appropriately.

MEASURE

The **KPI Library** provides a categorized list of potential indicators across clinical, operational, experience, equity, and financial domains. Select only those KPIs that are feasible and relevant. Indicators are tagged by evaluation tier to support low-capacity settings.

Evaluation is not a one-time exercise. Performance, workflow impact, and safety signals should be revisited as implementation matures or context changes.

At an appropriate interval after implementation, organizations should review whether the originally intended outcomes were achieved at a meaningful threshold, what evidence supports that conclusion, and what uncertainties remain. This retrospective review can inform sustainment, refinement, procurement, and future scaling decisions.

APPLY, INTERPRET and DECIDE

The **ROI Calculator** supports estimation of time savings, avoided events, and cost signals. Clearly define perspective, comparator, and time horizon before interpreting results. **Case Examples** may be used to support interpretation of early results and provide practical reference points for similar implementations.

The outputs of this toolkit are intended to support decision-making, not just measurement. At this stage, consider adoption, risk signals, and value together to determine the most appropriate next step. Reference the **Decision Guidance** section for support.

Interpretation Assumptions + Limitations

Metrics included in this toolkit are intended to support structured, proportional evaluation of clinical AI implementations. They support structured interpretation of impact in context. Where possible, intended outcomes and success thresholds should be specified before implementation to support meaningful comparison after go-live.

Across most implementations, interpretation of results relies on several common assumptions:

- Baseline and post-implementation periods are reasonably comparable.
- Workflow conditions remain relatively stable during measurement.
- Usage data and system logs accurately capture user activity.
- Self-reported measures reflect genuine experience.
- Time savings translate into usable clinical or operational capacity.
- Cost inputs include relevant one-time and recurring components.
- Observed changes are interpreted directionally, recognizing that attribution to AI may be partial.

Organizations should interpret results in context and document key uncertainties, particularly during early pilots or low-data implementations. This toolkit supports structured learning and decision-making, not definitive proof of impact.

Evaluation Process and Methods

Evaluation of clinical AI extends beyond performance metrics and financial return. Responsible use requires attention to safety, equity, transparency, sustainability, and long-term stewardship. These considerations are integrated throughout the toolkit and should be revisited over time as implementation matures.
This section outlines the evaluation process and the methods used within it, including the Minimum Viable Evaluation (MVE).

Minimum Viable Evaluation

(Basic/Low-Data)

This evaluation provides the simplest consistent approach for early evaluation. It allows any organization, even those with limited analytic capacity, to capture a basic before and after assessment and generate an early sense of impact.

Minimum Requirements (applies to all organizations)

The following elements should be included in every evaluation, regardless of organizational size or analytic capacity:

- Confirmation that evaluation activities use data already approved for clinical or operational use and do not introduce new data uses or privacy risks
- Identification of first- and second-order impacts relevant to the implementation
- A simple estimate of organizational effort and person-hours required for implementation and early use
- Basic model oversight and early feedback-data assessment to verify correctness and alignment with clinician judgment
- Completion of the minimum implementation-readiness checklist (workflow fit, data basics, staff comfort, time capacity, context factors, and simple evaluation feasibility)
- A brief environmental sustainability note where relevant (for example, computing intensity or carbon considerations)
- In most cases, a simple before-and-after comparison using two to four indicators and a brief narrative summary is sufficient for early-stage evaluation.

These requirements ensure consistency across organizations while keeping the Minimum Viable Evaluation achievable for small and low-data settings.

STEP 01

Define the AI Tool and Intended Impact

- One-sentence description of the AI tool
- Primary use case category (of the six defined categories)
- Expected impact in plain language

STEP 02

Capture a Simple Baseline (1 to 3 months)

- Use only data already available and already approved for this purpose
- Counts, turnaround times, clinician burden ratings
- Small or manual samples acceptable

STEP 03

Document Go-Live

- Implementation date
- Any contextual factors

STEP 04

Post-Implementation Measurement (1 to 3 months)

- Repeat the same simple indicators
- Manual audits allowed

STEP 05

Clinician Experience Check

- Three to five simple questions on workload, usability, trust, and perceived impact

STEP 06

Basic Model Oversight Check

- Review a small number of representative cases to confirm that AI outputs appear correct and appropriate. Document any clinician overrides, concerns, or unexpected outputs raised during early use.

STEP 07

Basic Before and After Comparison

- One small table
- Two to four indicators
- Short narrative summary

STEP 08

Narrative ROI Summary

- Key changes
- Efficiency or avoided-cost signals
- Clinician experience insights

STEP 09

Minimum ROI Signals

- Simple indicators of time saved or avoided events
- Direction of early ROI (positive, neutral, negative, uncertain)
- Brief narrative interpretation

• OPTIONAL •

Expanded Evaluation

(for more mature organizations)

The Expanded Evaluation builds directly on the Minimum Viable Evaluation and is optional for organizations with greater analytic capacity or higher-risk implementations.

Expanded does not mean expected. The Expanded Evaluation is optional and should be used only where organizational capacity, implementation risk, or decision stakes warrant it. A well-executed Minimum Viable Evaluation remains a valid and appropriate approach for many organizations.

Organizations with greater analytic capacity, richer data environments, or more complex implementations may choose to complete additional analyses. These optional components build on the Minimum Viable Evaluation and provide deeper insight into performance, context, and financial return.

These analyses strengthen confidence but remain context-dependent and interpretation-based.

Additional Baseline Options

- Longer baseline periods (6 to 12 months) to assess pre-intervention trends
- Monthly volume tables to understand workflow patterns over time
- Context and seasonal variation notes
- Ability to define comparator types (historical, parallel group, staggered rollout)

Expanded KPI Set

- Selection of additional clinical, operational, experience, equity, or financial KPIs
- Ability to include indicators beyond those captured in the Minimum Viable Evaluation
- Expanded baseline measurement for each KPI
- Flexibility to reflect organization-specific priorities

Enhanced Post-Implementation Measurement

- Repeat measurement of expanded KPIs
- Support for manual audits or time-motion studies
- Additional narrative notes to interpret early results
- Space to document implementation nuances influencing variation

Scenario Modeling (Low / Medium / High)

Scenario assumptions are examples. Teams should define the conditions that represent low, medium, and high impact in their context.

- Define expected outcomes under different adoption or utilization scenarios
- Adjust inputs such as volume, accuracy, or workload impact
- Use for planning, forecasting, or interpreting why observed results differ from expectations

Sensitivity Analysis

Variables in this section are placeholders. Teams should specify the factors most relevant to their implementation.

- Identify variables most likely to influence outcomes (for example, time saved, event rate changes, baseline uncertainty)
- Run low/high estimates to understand robustness of findings
- Helps teams interpret uncertainty in early evaluations

Financial Impact (Optional)

See full ROI calculator optional for advanced users.

- Estimate implementation effort and ongoing workload implications
- Include licensing or subscription costs
- Include avoided-event or time-savings benefits
- Generate simple net-benefit and ROI calculations

Comparator Summary

- Define comparator approach (parallel group, matched historical, staggered)
- Capture comparator KPI values before and after implementation
- Provide notes on contextual differences influencing comparability

Trend Analysis

- Twelve months of KPI values (baseline and post)
- Identify seasonal or contextual patterns
- Support dashboard-style review of adoption and impact trends
- Where repeated time points exist and randomization is not feasible, teams may consider quasi-experimental approaches such as interrupted time series to strengthen causal interpretation of observed trends

Expanded Narrative Summary

- High-level interpretation integrating all expanded components
- Clarification of what drove the observed results
- Identification of remaining uncertainties
- Recommendations and next steps for scaling, improving, or refining the AI tool

• OPTIONAL •

Responsible and Sustainable AI Considerations

(applies across all elements of the framework)

For many software-as-a-service tools, a formal sustainability assessment may not be required. In these cases, teams may simply confirm that vendor infrastructure, support, and licensing arrangements are sufficient for continued use.

01 Safety and Human Oversight

AI outputs should be subject to appropriate human review during early implementation. Evaluation does not replace clinical judgment or regulatory requirements. Oversight mechanisms should remain in place beyond initial rollout. In some clinical AI applications, the ability for users to understand the basis of AI-generated outputs (often referred to as explainability) can influence clinician trust, appropriate use, and safe decision-making.

02 Equity and Fairness

Organizations should monitor for differential performance or unintended impacts across patient populations. Where feasible, equity signals should be incorporated into evaluation and revisited as data accumulates.

03 Transparency and Communication

Clinicians and staff should understand what the AI tool does, what it does not do, and how to report concerns. Clear communication supports trust and responsible adoption.

04 Environmental and Infrastructure Considerations

Where relevant, organizations may note compute intensity, infrastructure requirements, or sustainability implications. Not all implementations require formal environmental assessment but awareness is encouraged.

05 Long-Term Stewardship

Performance may change over time due to workflow shifts, population changes, or model updates. Periodic reassessment supports sustained value and reduces drift risk.

06 Change Management and User Readiness

Successful implementation also depends on change management, digital literacy, and appropriate user education, including clinician, staff, and where relevant, patient-facing communication. This toolkit acknowledges these as enabling conditions but does not provide a full implementation or training playbook.

• OPTIONAL •

Simple Evaluation Techniques

This section provides a set of simple, practical techniques that any organization can use to strengthen interpretation of early results. These techniques help teams forecast expected outcomes, choose appropriate comparators, assess attribution, and identify unintended consequences in a low-data or early-implementation environment.

Anchoring each section with one or two simple, concrete use cases, whether anonymized or hypothetical, will help teams translate the concepts into practice and understand how to apply the framework in real-world settings. This toolkit includes a few brief use cases to illustrate how the concepts can be applied.

Forecasting

- Use recent trends to project forward
- Provide low, medium, and high scenarios

Attribution

- Consider parallel workflow changes
- Consider common confounders such as staffing changes, patient volume fluctuations, or concurrent improvement initiatives
- Seek clinician confirmation
- Use contribution logic rather than strict causality

Comparators

- Before and after
- Historical baseline
- Matched unit
- Staggered rollout

Unintended Consequences

- Workflow disruption
- Misclassification concerns
- Alert fatigue
- Equity implications

• OPTIONAL •

AI Type Considerations

The evaluation approach in this toolkit applies across AI implementations. However, different AI types may warrant emphasis on different risks, performance indicators, and oversight considerations. The following table highlights common areas of nuance. This section does not introduce new requirements.

AI TYPE	TYPICAL USE	EVALUATION EMPHASIS	OVERSIGHT CONSIDERATIONS	KPI SENSITIVITIES
Predictive Risk Models	Stratification, risk scoring, early detection	Accuracy, calibration, downstream impact, variability reduction	Drift monitoring, bias across populations	Accuracy, Error Rate, Intended Population Coverage, Variability Reduction
Clinical Decision Support Alerts	Real-time recommendations during care	Adoption, override rate, workflow fit	Alert fatigue, appropriateness of use	User Adoption Rate, AI Recommendation Override Rate, Clinical Workflow Fit
Generative Documentation Tools	Drafting notes, summaries, communication	Time saved, text quality, clinician trust	Hallucination risk, instructions followed	Hours Saved, Text Quality, Coherence/ fluency, Clinician-Reported Clinical Value
Workflow Automation Tools	Task routing, triage, scheduling	Efficiency, task completion time, system reliability	Process errors, system downtime	Task Completion Time, System Availability, Error Rate
Diagnostic / Imaging AI	Image interpretation, pattern detection	Sensitivity/ specificity, safety signals	Model performance drift, false positives/ negatives	Accuracy, Observed Harm or Safety Concerns, Error Rate

PREPARE

Implementation Readiness Checklist

The checklist on the following page outlines the basic conditions that support successful adoption and practical evaluation of an AI tool. It is designed to help organizations of all sizes, including small clinics with limited analytic capacity, determine whether they have the essential groundwork in place before beginning measurement. It ensures that workflow fit, data basics, staff comfort, and simple feasibility have been considered.

- [ ] **Workflow Fit**
The team understands how the AI tool fits into the daily workflow, who will use it, and at what point in the patient encounter it is accessed. A few test cases have been tried to ensure the tool does not add friction or unnecessary steps. Have the workflow changes required for the AI tool to deliver its intended benefit been identified, documented, and communicated to affected staff?

- [ ] **Data Basics**
The organization has the minimum data required for the tool to function and can capture a simple baseline for one or two indicators using existing data sources. This assumes that existing organizational privacy, consent, and data-use approvals apply and that no new data uses are introduced for evaluation purposes.

- [ ] **Clinician and Staff Comfort**
Clinicians and staff understand what the tool does and does not do and know how to note issues, unexpected outputs, or early observations during initial use.

- [ ] **Time and Capacity**
A short period of protected time exists to monitor early use and reflect on impacts. The team can assess whether the tool appears to save time or effort relative to the change required to adopt it.

- [ ] **Context Factors**
Recent changes in schedules, staffing levels, patient volumes, or service patterns are documented, as these factors may influence evaluation results during the baseline or early measurement period.

- [ ] **Equity and Experience**
The tool appears to work consistently across the typical patient population served by the organization and does not hinder clinician or staff workflow experience.

- [ ] **Evaluation Feasibility**
The organization has selected one or two practical indicators to measure, identified a simple before–after assessment period, and determined who will summarize results in a brief narrative.

- [ ] **Change Management and User Readiness**
The organization has identified what orientation, training, communication, and support will be needed for clinicians, staff, and where relevant, patients or other end users during early implementation.

MEASURE

KPI Library

The KPI Library is a structured set of metrics organizations can select from when designing an AI evaluation approach.

The toolkit includes a starter KPI Library that provides commonly used measures across five domains. These KPIs support both the Minimum Viable Evaluation and the Expanded Evaluation. Organizations can select from this list or add their own indicators if needed.

KPI categories included in the library:

- Clinical
- Operational
- Experience (clinician or patient)
- Equity
- Financial

These categories reflect the domains most used in evaluating the impact of clinical AI tools and are used consistently across the toolkit and supporting tools.

Organizing KPIs in this way provides a consistent structure for measurement while allowing flexibility across different implementation settings.

Within each category, the library includes example indicators that represent typical outcomes, process measures, or experience measures used in AI evaluation. This is a starter library, and the list can expand and refine over time to reflect additional use cases, population contexts, and the needs of smaller or lower-data organizations.

Metric, KPI, ROI Input Name	Metric Category	Common Across Most AI Types?	Evaluation Tier	Primary Domain	Secondary Lens (optional)	Plain language definition
Accuracy	Techical/Model Performance	Yes	Minimum (Low-Data Feasible)	Clinical		The proportion of AI outputs that are correct when compared to a reference standard.
AI Recommendation Override Rate	User Experience	Yes	Minimum (Low-Data Feasible)	Operational	Safety	How often users choose not to follow AI recommendations, indicating trust, fit, or concern.
Click Through Rate (CTR)	Implementation Evaluation	Context-specific	Advanced (High Analytic Capacity)	Operational		Represents the percentage of people who take action after an event
Clinical Use Appropriateness	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Clinical	Safety	The extent to which the AI tool is used for the clinical purposes and patient groups it was intended for.
Clinical Workflow Fit	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Operational		How well the AI fits into existing clinical workflows without disrupting care delivery.
Clinician-Reported Clinical Value	User Experience	Yes	Minimum (Low-Data Feasible)	Experience (clinician or patient)	Clinical	Clinicians' assessment of whether the AI meaningfully supports clinical decision-making or care delivery.
Coherence / Fluency	Techical/Model Performance	Context-specific	Advanced (High Analytic Capacity)	Experience (clinician or patient)	Quality	Clarity and logical order of a response.
Contextual Suitability	Implementation Evaluation		Expanded (Moderate Capacity)	Operational	Equity	How well the AI works across different care contexts, resource levels, or local practices.
Cost of Implementation	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Financial	Operational	The total cost required to introduce the AI tool, including licensing, setup, and staff time.
Downstream Benefits	Implementation Evaluation	Yes	Expanded (Moderate Capacity)	Financial	Operational	Secondary or indirect benefits that occur as a result of initial improvements enabled by an AI solution.
Ease of Use	User Experience	Yes	Minimum (Low-Data Feasible)	Experience (clinician or patient)	Trust	How easy and intuitive the AI tool is for users to operate within their normal workflow.
Error Rate	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Clinical	Safety	The frequency with which the AI produces incorrect outputs relative to a defined reference standard.
Hours Saved	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Operational	Financial	Represents aggregate time savings across the organization. The total staff time saved across a defined period due to AI-assisted or automated tasks.
Implementation Effort	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Operational	Financial	The amount of staff time and organizational effort required to implement and maintain the AI tool.
Intended Population Coverage	Implementation Evaluation		Expanded (Moderate Capacity)	Equity	Operational	The extent to which the AI reaches the populations and settings it was designed to support.
Model Latency	Techical/Model Performance	Context-specific	Advanced (High Analytic Capacity)	Operational	Experience	The time required for the AI system to generate a response after receiving a request.
Observed Harm or Safety Concerns	Implementation Evaluation	Yes	Minimum (Low-Data Feasible)	Clinical	Safety	The presence and frequency of reported patient safety concerns or unintended negative effects related to AI use.

DOWNLOAD THE KPI LIBRARY

APPLY, INTERPRET, AND DECIDE

ROI Calculator

The ROI calculator provides both a simple and an expanded approach for estimating value. It supports early directional ROI signals as well as full financial calculations for organizations with greater analytic capacity.

Basic ROI (Minimum Viable Evaluation)

Time saved, Avoided events, Directional ROI classification

Full ROI (Expanded Evaluation)

Licensing cost, Implementation effort, Ongoing costs, Avoided events, Time value, Net benefit, ROI %, Payback period

Before calculating full ROI, organizations should clearly define the perspective (who realizes the value), the comparator (current state), and the time horizon over which value is assessed. Failure to define these elements can lead to inconsistent or misleading ROI conclusions.

These inputs should reflect full lifecycle costs, including one-time implementation effort, recurring operating costs, and any additional costs associated with scaling or sustained use. Not all value needs to be monetized; impacts related to quality, experience, equity, or safety may be captured directionally or narratively alongside financial measures. Where precise data are not available, inputs may be expressed as reasonable ranges and key assumptions should be noted explicitly.

For Canadian organizations preparing more formal value cases, advanced ROI approaches should align with Canadian health technology assessment norms and expectations. In practice, these are reflected in the guidance commonly used by organizations such as CADTH and provincial health technology assessment programs and are referenced here for alignment rather than compliance.

For organizations preparing formal business cases, system-level approvals, or multi-year investment decisions, a separate Advanced ROI Reference is available. This reference builds on the concepts above and aligns with international digital health and AI economic evaluation frameworks. It is optional and intended for organizations with greater analytic and resourcing capacity.

DOWNLOAD THE ROI CALCULATOR

APPLY, INTERPRET, AND DECIDE

Decision Guidance

The outputs of this toolkit are intended to support decision-making, not just measurement. At this stage, consider adoption, risk signals, and value together to determine the most appropriate next step.

Continue / Scale

Proceed with broader implementation or expansion when:

- Adoption is meaningful and showing signs of growth
- No significant unresolved risks are identified (for example, quality, safety, workflow, or equity concerns)
- Directional ROI is positive and supported by early signals
- Continue with Conditions

Proceed with targeted adjustments and ongoing monitoring when:

- Adoption is uneven or concentrated within a subset of users
- Specific risks or limitations are identified (for example, content quality, editing burden, workflow fit, or equity considerations)
- Directional ROI is positive but dependent on key assumptions or incomplete data

Focus on addressing identified gaps before scaling further.

Pause / Reassess

Pause further rollout and reassess when:

- Adoption is low or declining
- Significant risks or concerns remain unresolved
- There is no clear or consistent signal of value

Additional data collection, workflow adjustment, or tool refinement may be required before proceeding.

Important considerations

- These are guides, not thresholds or rules
- Decisions should not rely on a single input (for example, ROI alone)
- Early findings are directional and should be revisited as more data becomes available
- Ongoing oversight is expected, with reassessment at defined intervals (for example, 3 to 6 months)

ADDITIONAL RESOURCES

Case Examples

These case example cards illustrate how evaluation concepts and value interpretation may be applied in different clinical and operational contexts.

The examples are intentionally illustrative and directional. They are not validated case studies and do not represent formal outcome evaluations. Their purpose is to demonstrate how organizations might approach early assessment, interpret signals of impact, and identify contextual factors that influence results.

Each example also highlights contextual factors such as workflow integration, user training, or operational readiness that can influence early evaluation results.

These examples can be used alongside the MVE, KPI selection, and ROI interpretation steps to guide application.

Emergency Department Triage Risk Identification

PURPOSE

Illustrative example showing how early-stage evaluation can focus on safety, workflow impact, and feasibility rather than full outcome validation.

DOMAINS

Primary: Clinical

Secondary: Operational

CONTEXT / PROBLEM

Emergency departments manage high patient volumes under time pressure. Triage nurses document symptoms in free-text notes, and subtle indicators of higher risk may be difficult to identify consistently during peak periods.

METRICS / INDICATORS USED

- ✔ Proportion of high-acuity cases flagged by the tool
- ✔ Time from triage to clinician assessment
- ✔ Clinician feedback on alert usefulness
- ✔ Volume of alerts generated per shift

OBSERVED VALUE DIMENSIONS

Quality: earlier clinical assessment for higher-risk patients

Time: reduced delays to provider review

AI TOOL OR CAPABILITY

A natural language processing tool reviews triage notes in real time and flags encounters that may warrant earlier clinical review based on symptom patterns.

OBSERVED OR REPORTED IMPACT

Early pilots suggested faster clinician review for some higher-risk patients and may have improved confidence that critical symptoms were not overlooked. Staff noted an increase in alerts during peak hours, requiring monitoring for potential alert fatigue.

NOTES / CONTEXTUAL FACTORS

Evaluation focused on directional trends and workflow impact rather than patient outcomes. Small sample sizes and variable documentation practices limited precision and attribution. Brief orientation sessions were provided to triage nurses to explain how alerts should be interpreted and when clinical judgment should override tool recommendations.

Ambient Clinical Documentation Assistant

PURPOSE

Illustrates evaluation of low clinical risk AI with high operational and experience impact.

CONTEXT / PROBLEM

Clinicians report spending significant time on documentation, often outside clinical hours, contributing to burnout and reduced patient interaction.

OBSERVED VALUE DIMENSIONS

Experience: improved clinician satisfaction

Time: reduced documentation burden

OBSERVED OR REPORTED IMPACT

Clinicians reported reduced documentation time and improved focus during patient visits. Early evaluation suggested variability in note quality across visit types.

DOMAINS

Primary: Operational

Secondary: Experience

METRICS / INDICATORS USED

- ✔ Average documentation time per visit
- ✔ After-hours charting time
- ✔ Clinician satisfaction and perceived burden
- ✔ Frequency of manual edits required

AI TOOL OR CAPABILITY

An ambient AI tool captures clinical conversations and generates draft visit notes for clinician review and approval within the electronic health record.

NOTES / CONTEXTUAL FACTORS

Evaluation emphasized usability and workflow fit. Formal cost savings were not calculated during the pilot phase, and results were interpreted directionally. Early implementation included short onboarding sessions to help clinicians understand how to review, edit, and approve AI-generated notes within existing documentation workflows.

Inpatient Deterioration Prediction Tool

PURPOSE

Illustrates higher-risk AI requiring more structured evaluation and monitoring.

CONTEXT / PROBLEM

Patient deterioration on inpatient units may go unrecognized between routine vital sign checks, leading to delayed intervention.

OBSERVED VALUE DIMENSIONS

Experience: earlier clinical intervention

Time: faster escalation of care

OBSERVED OR REPORTED IMPACT

Initial deployment appeared to increase early identification of some deteriorating patients but also generated a high volume of alerts, requiring threshold adjustments and ongoing review.

DOMAINS

Primary: Clinical

Secondary: Operational

METRICS / INDICATORS USED

- ✔ Sensitivity for detecting deterioration events
- ✔ Alert volume and response rates
- ✔ Rapid response team activations
- ✔ Clinician-reported alert usefulness

AI TOOL OR CAPABILITY

A predictive model uses vital signs, laboratory results, and clinical data to estimate risk of deterioration and generate alerts for care teams.

NOTES / CONTEXTUAL FACTORS

Ongoing monitoring was required to balance potential safety benefits with alert fatigue. Attribution of outcomes to the AI tool alone was challenging in a complex inpatient environment. Implementation included guidance for nursing and medical staff on interpreting alerts and incorporating them into existing escalation protocols and rapid response workflows.

AI-Supported Medication Refill Prioritization

PURPOSE

Illustrates low-risk AI that may improve workflow efficiency, adherence, and safety.

CONTEXT / PROBLEM

Primary care clinics face excessive medication refill requests, which can lead to delays for some patients and disproportionate staff workload.

OBSERVED VALUE DIMENSIONS

Experience: improved patient access

Time: reduced administrative effort

Equity: more consistent prioritization across patients

OBSERVED OR REPORTED IMPACT

Staff reported smoother handling of refill queues and faster turnaround for higher-risk medications. Patient complaints related to delays appeared to decrease during early use.

DOMAINS

Primary: Experience

Secondary: Operational

METRICS / INDICATORS USED

- ✔ Time to refill prescription
- ✔ Staff time spent triaging requests
- ✔ Volume of urgent escalations
- ✔ Clinician satisfaction on obtaining timely refill

AI TOOL OR CAPABILITY

An AI tool handles routine refills and prioritizes refill requests based on medication type, patient history, and urgency indicators such as chronic condition management.

NOTES / CONTEXTUAL FACTORS

Improvements were observed in workflow handling, though staffing changes limited attribution. Evaluation relied on pre/post comparison and staff feedback and was interpreted directionally. Staff received brief guidance on how automated prioritization worked and when manual review or override was appropriate.

AI-Supported Documentation Assistant in Primary Care

PURPOSE

Illustrates a low-risk AI tool where evaluation focuses on workflow efficiency and clinician experience.

CONTEXT / PROBLEM

Primary care clinicians often spend significant time documenting visits in the electronic medical record after patient encounters, contributing to administrative burden and reduced time available for patient care.

OBSERVED VALUE DIMENSIONS

Time: reduced documentation time

Experience: reduced administrative burden for clinicians

OBSERVED OR REPORTED IMPACT

Clinicians reported modest reductions in documentation time and fewer notes completed after clinic hours. Some variation in note quality required clinicians to review and edit drafts carefully, particularly for complex visits.

DOMAINS

Primary: Operational

Secondary: Experience

METRICS / INDICATORS USED

- ✔ Average documentation time per visit
- ✔ Proportion of notes requiring substantial edits
- ✔ Clinician satisfaction with note quality
- ✔ Time spent completing documentation after clinic hours

AI TOOL OR CAPABILITY

An AI documentation assistant listens to the clinical conversation and generates a draft progress note for the clinician to review and edit before finalizing in the electronic medical record.

NOTES / CONTEXTUAL FACTORS

Evaluation focused on workflow impact and clinician experience rather than clinical outcomes. The clinic conducted a short Minimum Viable Evaluation using time-tracking and clinician feedback. Results were interpreted cautiously because documentation patterns varied across clinicians and visit types. Implementation guidance emphasized clinician review of all AI-generated notes before finalization.

ADDITIONAL RESOURCES

Standardized Taxonomy

AI Clinical Function Use Case Categories

Triage & Risk Stratification
AI tools for early detection of high-risk patients, symptom assessment, and predictive analytics.

Diagnosis & Decision Support
AI-driven technologies supporting radiology, pathology, and clinical decision-making.

Treatment & Intervention
AI-enhanced robotic surgery, personalized medicine algorithms, and real-time intervention tools.

Follow-up & Remote Monitoring
AI-powered patient monitoring, wearable device analytics, and predictive analytics for post-discharge care.

Patient Education & Engagement
AI-driven chatbots, virtual assistants, and personalized patient education platforms.

Communication & Coordination
AI solutions for clinical workflow automation, provider communication, and administrative efficiency.

Stage of Deployment

Planning
Announced or being designed; not yet piloted.

Pilot
Limited-scale testing in a real-world setting.

In Use
Deployed in routine practice but not yet scaled across multiple sites or full systems.

Scaled
Broad deployment across multiple units/sites within an organization.

System-wide
Adopted at the full health-system or multi-institutional level.

Clinical Venues

Hospital
Inpatient, general or specialized hospital facilities (regional, community, academic/tertiary).

Acute Care
Units/services for time-sensitive, high-intensity medical needs (ED, ICU, critical care, oncology wards, diagnostic imaging within hospitals).

Ambulatory Care Centre
Outpatient hospital-affiliated clinics (e.g., wound, diabetes, oncology follow-up).

Community Health
Non-hospital local health centres, community clinics, Indigenous health centres.

Primary Care
Family practice, GP clinics, virtual-first care platforms.

Long Term Care
Nursing homes, residential facilities, supportive housing for seniors.

Public Health
Population-level health monitoring, screening, disease surveillance.

Specialty Clinic
Stand-alone or hospital-affiliated specialty units (endoscopy, fertility, orthopedic).

Health System
Multi-site or integrated health networks spanning hospitals, clinics, LTC, and/or public health.

Diagnostic Imaging
Radiology and imaging services, including standalone centres or hospital-based units (e.g., MRI, CT, ultrasound, mammography, stroke/neuroimaging).

Technology Types

AI (general)
Used when details of the underlying model are not specified or span multiple techniques.

Machine Learning (ML)
Algorithmic models trained on structured or semi-structured data for predictions or classifications.

Deep Learning (DL)
Neural network–based AI, typically used for imaging, signal, or complex pattern recognition.

Natural Language Processing (NLP)
AI tools for analyzing and interpreting human language (transcription, summarization, entity extraction).

Large Language Models (LLM)
Transformer-based generative models (e.g., GPT-type), specialized for text generation and reasoning.

Computer Vision
AI applied to image/video recognition, detection, and segmentation.

Robotics
AI-enabled robotic systems in surgical, diagnostic, or rehabilitation contexts.

Source: https://digitalhealthcanada.com/membership/chief-executive-forum/inititatives/ai-in-action

ADDITIONAL RESOURCES

Glossary

These **plain-language term definitions** are provided as supporting materials for the Clinical AI Evaluation Toolkit. They are intended to promote consistent interpretation of key evaluation and ROI concepts across organizations. The definitions are practical and directional. They are not formal technical or academic definitions.

Alert Burden

Plain language definition:
The number of alerts generated by an AI system and the effort required by staff to review and respond to them.

Why this term matters:
High alert burden can lead to alert fatigue, reducing trust in the tool and potentially undermining safety benefits.

Example or clarification:
A predictive monitoring tool may initially generate many alerts, requiring threshold adjustments to balance early detection with a manageable alert volume.

Related terms:
- Workflow Impact
- Ongoing Monitoring
- Usability

Attribution

Plain language definition:
The extent to which observed changes can reasonably be linked to the AI tool rather than to other changes happening at the same time.

Why this term matters:
Healthcare environments are complex, and improvements often occur alongside staffing, workflow, or policy changes.

Example or clarification:
If staffing levels change during an AI pilot, it may be difficult to determine how much improvement is due to the tool alone.

Related terms:
- Pre/Post Comparison
- Directional Impact
- Early-Stage Evaluation

Change Management

Plain Language Definition

The structured process used to support individuals and teams in adopting new tools, workflows, or ways of working introduced by an AI implementation.

Why This Matters

Even well-designed AI tools can fail to deliver value if users are not prepared, supported, or confident in how the tool fits into their workflow. Change management helps ensure that clinicians, staff, and other users understand the purpose of the technology, receive appropriate training, and have channels to provide feedback or raise concerns. Without deliberate attention to change management, adoption may be inconsistent, benefits may not materialize, and evaluation results may be misleading.

Example or Clarification

Before introducing an AI triage tool in the emergency department, the implementation team provides staff orientation sessions, integrates the tool into existing clinical protocols, and establishes a feedback channel for clinicians to report usability issues or unexpected outputs.

Related Terms

- User readiness
- Implementation effort
- Workflow impact
- Usability

Directional Impact

Plain language definition:

Evidence that outcomes are moving in a positive or negative direction after an AI tool is introduced, without claiming precise or final results.

Why this term matters:

Early pilots often lack large sample sizes or control groups. Directional impact allows teams to learn whether a tool appears helpful before committing to more rigorous evaluation.

Example or clarification:

Clinicians reporting faster review of higher-risk patients suggests a positive directional impact, even if patient outcomes are not formally measured.

Related terms:

- Pre/Post Comparison
- Proxy Measure
- Early-Stage Evaluation

Early-Stage Evaluation

Plain language definition:

A lightweight approach to assessing an AI tool during a pilot or early rollout that focuses on safety, usability, and workflow impact rather than final outcomes or full financial return.

Why this term matters:

Most clinical AI tools are first tested in real-world settings, where full validation or cost analysis is not feasible. Early-stage evaluation helps teams decide whether a tool is safe and useful enough to continue.

Example or clarification:

Early evaluation may focus on whether AI-generated alerts are helpful to clinicians and whether the tool reduces delays in clinical review, without attempting to measure long-term patient outcomes or cost savings.

Related terms:

- Workflow Impact
- Feasibility
- Pre/Post Comparison

Evaluation Depth

Plain Language Definition

The level of detail, rigor, and breadth used to assess the performance, impact, and value of an AI implementation.

Why This Matters

Not all AI implementations require the same level of evaluation. Early pilots, lower-risk tools, or organizations with limited analytic capacity may rely on a Minimum Viable Evaluation, which focuses on a small set of meaningful indicators. Higher-stakes implementations, system-wide deployments, or initiatives with stronger evidence requirements may require a deeper or more comprehensive evaluation approach. Recognizing different levels of evaluation depth helps organizations apply the toolkit pragmatically rather than assuming every implementation must meet the same evaluation standard.

Example or Clarification

A hospital testing an AI documentation assistant during a short pilot may track a small number of indicators such as clinician time savings and satisfaction. This represents a lower evaluation depth appropriate to an early-stage test. A health system deploying an AI triage model across multiple emergency departments may conduct a deeper evaluation that includes clinical outcomes, workflow impact, safety monitoring, and cost analysis.

Related Terms

- Minimum Viable Evaluation
- Early-Stage Evaluation
- Structured Outcomes Review
- Pre/post comparison

Evaluation Domains

Plain language definition:

The five categories used to organize indicators in the toolkit: Clinical, Operational, Experience (clinician or patient), Equity, and Financial.

Why this term matters:

Organizing evaluation using consistent domains helps ensure that value is assessed comprehensively and proportionately. Different AI tools may emphasize different domains, but all evaluations should consider which domains are most relevant.

Example or clarification:

A documentation assistant may primarily influence operational and experience domains, while a deterioration prediction tool may emphasize clinical and safety domains.

Related terms:

- First-Order Impact
- Second-Order Impact
- Indirect ROI

Experience Impact

Plain language definition:

The effect an AI tool has on the experience of patients or staff, including satisfaction, stress, and perceived workload.

Why this term matters:

Improvements in experience can support adoption, sustainability, and overall quality of care.

Example or clarification:

Reduced after-hours charting can improve clinician experience even if visit length stays the same.

Related terms:

- First-Order Impact
- Second-Order Impact
- Indirect ROI
- Workflow Impact

Feasibility

Plain language definition:
Whether an AI tool is practical and can be successfully implemented and used in real-world settings, considering staffing, data quality, and operational constraints.

Why this term matters:
A tool can perform well in theory, but may be difficult to sustain in practice due to documentation variability, staffing pressures, or workflow misalignment.

Example or clarification:
In settings where clinical documentation practices vary widely, an AI tool may work well for some users, yet inconsistently for others.

Related terms:
- Early-Stage Evaluation
- Workflow Impact
- Usability

First-Order Impact

Plain language definition:
The immediate and direct change associated with using an AI tool, such as time saved, alerts generated, or tasks avoided.

Why this term matters:
Early evaluation often focuses on first-order impacts because they are easier to observe and measure than longer-term outcomes.

Example or clarification:
Reduced documentation time per visit is a first-order impact of an ambient documentation tool.

Related terms:
- Second-Order Impact
- Directional Impact
- Indirect ROI

Implementation Effort

Plain language definition:
The time and organizational resources required to introduce, monitor, and support an AI tool during early use.

Why this term matters:
Understanding effort helps interpret early ROI signals. A tool that saves time but requires significant monitoring or workflow change may have different value implications.

Example or clarification:
Implementation effort may include clinician training time, configuration work, workflow adjustments, or early monitoring activities.

Related terms:
- Indirect ROI
- Minimum ROI Signal
- Attribution

Indirect ROI

Plain language definition:
Benefits created by an AI tool that do not immediately translate into direct cost savings but still contribute to organizational value.

Why this term matters:
Focusing only on direct financial return can overlook meaningful improvements in efficiency, safety, or staff well-being.

Example or clarification:
Reduced administrative effort and avoided delays represent indirect ROI even when no formal cost savings are calculated.

Related terms:
- Second-Order Impact
- Proxy Measure
- Experience Impact

Minimum ROI Signal

Plain language definition:
An early directional indication of whether an AI tool appears to generate value, based on limited data and short-term observation.

Why this term matters:
In early pilots, it may not be possible to calculate full financial return. A minimum ROI signal helps teams decide whether value appears positive, neutral, negative, or uncertain.

Example or clarification:
If time savings are observed and no safety concerns arise, the minimum ROI signal may be considered positive, even if precise cost savings are not calculated.

Related terms:
- Directional Impact
- Pre/Post Comparison
- Implementation Effort

Minimum Viable Evaluation

Plain language definition:
The simplest consistent approach to evaluating a clinical AI tool using a small set of before-and-after indicators and brief narrative interpretation.

Why this term matters:
Not every organization has the capacity for complex analysis. The Minimum Viable Evaluation allows teams to generate an early, structured signal of impact using existing data and limited time.

Example or clarification:
An organization may compare one or two indicators, such as turnaround time or clinician workload, before and after implementation and summarize the direction of change.

Related terms:
- Pre/Post Comparison
- Minimum ROI Signal
- Implementation Effort

Ongoing Monitoring

Plain language definition:
Continued review of performance, alerts, and unintended effects after an AI tool is deployed, particularly during early and Expanded Evaluation phases.

Why this term matters:
Some AI tools require adjustment over time to remain safe, effective, and aligned with clinical workflows.

Example or clarification:
A deterioration prediction tool may require regular threshold tuning to manage alert fatigue while maintaining safety benefits.

Related terms:
- Alert Burden
- Feasibility
- Workflow Impact

Pre/Post Comparison

Plain language definition:
An evaluation approach that compares outcomes before and after an AI tool is introduced using the same indicators, without using a separate control group.

Why this term matters:
Pre/post comparison is often the most feasible method for early pilots and low-capacity settings.

Example or clarification:
Teams may compare task completion times or staff workload before and after AI-supported medication refill prioritization is introduced.

Related terms:
- Early-Stage Evaluation
- Directional Impact
- Attribution

Proxy Measure

Plain language definition:
An indirect indicator used when the outcome of interest is difficult to measure directly.

Why this term matters:
Important outcomes such as clinician burden, access, or experience are often hard to quantify in early evaluations.

Example or clarification:
After-hours charting time may be used as a proxy for documentation burden when evaluating an ambient documentation tool.

Related terms:
- Directional Impact
- Pre/Post Comparison
- Indirect ROI

Second-Order Impact

Plain language definition:
The indirect or downstream effect that may occur as a result of first-order changes, such as improved access, reduced burnout, or fewer avoidable visits.

Why this term matters:
Some value from AI tools appears over time and may not be immediately measurable during early evaluation.

Example or clarification:
If documentation time decreases, clinicians may see more patients or experience less after-hours workload. These are second-order impacts.

Related terms:
- First-Order Impact
- Indirect ROI
- Directional Impact

Structured outcomes review

Plain language definition:
A deliberate review conducted after implementation to assess whether the intended outcomes of an AI tool were achieved, what evidence supports that conclusion, and what uncertainties or unintended effects remain.

Why This Matters
Early-stage AI implementations often rely on directional indicators rather than definitive proof of impact. A structured outcomes review helps organizations revisit the original goals of the implementation, assess whether meaningful signals of value or improvement have emerged, and decide whether to sustain, scale, refine, or discontinue the solution. It also supports institutional learning by documenting what worked, what did not, and why.

Example or Clarification
Six months after deploying an AI clinical documentation assistant, a hospital conducts a structured outcomes review. The team compares baseline and post-implementation metrics such as documentation time per encounter, clinician satisfaction, and note completeness. While some improvements are observed, the review also identifies workflow adjustments and additional training needs before wider rollout.

Related Terms
- Minimum Viable Evaluation
- Early-Stage Evaluation
- Ongoing monitoring
- Pre/post comparison

Usability

Plain language definition:
How easy an AI tool is for clinicians or staff to understand, interact with, and incorporate into their work.

Why this term matters:
Even effective tools may fail if they are confusing, time-consuming, or difficult to use in practice.

Example or clarification:
Clinician feedback on whether alerts are helpful or distracting provides insight into tool usability.

Related terms:
- Workflow Impact
- Feasibility
- Experience Impact

User Readiness

Plain language definition:
The degree to which clinicians, staff, or other intended users are prepared to effectively and safely use an AI tool in their day-to-day work.

Why This Matters
AI tools often introduce new information flows, decision supports, or documentation processes. If users do not understand how the tool works, when to rely on it, or how to interpret its outputs, the technology may be underused, misused, or resisted. Assessing user readiness helps organizations determine whether adequate training, guidance, and workflow alignment are in place before or during implementation.

Example or Clarification
A clinic preparing to introduce an AI risk stratification model ensures that physicians understand how risk scores are generated, what actions are recommended, and when clinical judgment should override model outputs.

Related Terms
- Change Management
- Implementation Effort
- Usability
- Workflow Impact

Workflow Impact

Plain language definition:
The way an AI tool affects how work is done day-to-day, including time spent on tasks, coordination effort, and efficiency within clinical or operational workflows.

Why this term matters:
AI tools that disrupt workflows or add burden are unlikely to be adopted, even if they perform well technically.

Example or clarification:
An AI documentation tool may reduce time spent charting after hours but still require adjustments to ensure notes are appropriate across different visit types.

Related terms:
- Feasibility
- Usability
- Experience Impact
- Workflow impact

About Digital Health Canada

Digital Health Canada connects, inspires, and empowers those enabling digital healthcare in Canada. Our members are a diverse community of accomplished, influential professionals working to make a difference in advancing healthcare through information technology. Digital Health Canada fosters network growth and connection; brings together ideas from multiple segments for incubation and advocacy; supports members through professional development at the individual and organizational level; and advocates for the Canadian digital health industry.

About CHIEF Executive Forum

Digital Health Canada's CHIEF Executive Forum provides a place for senior professionals and leaders in digital health and healthcare to collaborate, exchange best practices, address professional development needs, and offer their expertise in setting the agenda for the effective use of information and technology to improve health and healthcare in Canada. Members contribute their active participation, industry experience, and in-depth insight to working groups, publications, and health informatics discussions at the semi-annual CHIEF Symposia, and throughout the membership year.

Learn more at digitalhealthcanada.com