DRAFT REPORT

NATIONAL QUALITY FORUM
Driving measurable health improvements together

Quality Measures Using Artificial Intelligence (AI) Methods: Guidance and Recommendations for Developing, Selecting, and Implementing Measures in Accountability Programs

DRAFT REPORT
SEPTEMBER 8, 2025

ABOUT THE NATIONAL QUALITY FORUM

The National Quality Forum (NQF) is a not-for-profit, nonpartisan, membership-based organization that works to improve healthcare outcomes, safety, and affordability for all people. Our unique role is to bring all voices to our table to forge multistakeholder consensus on quality measurement and improvement standards and practices that achieve measurable health improvements for all. NQF is a proud affiliate of The Joint Commission. Learn more at www.qualityforum.org.

ABOUT THE GORDON AND BETTY MOORE FOUNDATION

This report is funded by the Gordon and Betty Moore Foundation.

Impact Statement

Purpose

To address the current gap in existing governance and guidance documents on the use of trustworthy artificial intelligence (AI) in quality measurement, National Quality Forum (NQF) produced this consensus-based report to provide guidance and recommendations for the development, selection, and implementation of AI-enabled quality measures in accountability programs. Accountability programs include accreditation, pay-for-performance, public reporting, and value-based payment programs.

Key Findings

The multistakeholder technical expert panel (TEP) convened by NQF for this report developed six strategies to advance the use of trustworthy AI-enabled quality measures in accountability programs, designed a five-step implementation roadmap, and identified four key actors involved in implementing the strategies. The TEP recommends that the quality measurement field employ the following six strategies, operationalized in the report, to advance the trustworthy use of AI-enabled quality measures:

1. Provide precise and transparent information about the AI-derived component (i.e., a component of a measure identified and/or calculated using AI methods), including data used in development and testing
2. Optimize performance of the AI-derived component through testing and tuning
3. Define the capabilities required to implement the AI-derived component (feasibility)
4. Assess regularly for unintended consequences, including bias
5. Prioritize ongoing monitoring and maintenance
6. Support an ecosystem that enables information sharing and feedback across key actors

The roadmap spans measure development and testing, selection, preparation for implementation, implementation across entities, and monitoring and maintenance. At each step, NQF and the TEP created actionable implementation roadmaps that detail the responsibilities and actions of each type of key actor—program owners (i.e., organizations responsible for administering national, regional, state, or local, public or private-sector accountability programs), measure developers, measured entities, and measure implementation vendors.

Applications

This guidance is intended primarily for accountability program owners because they are responsible for implementing quality measurement approaches designed to drive improvements in care. The secondary audience includes measure developers, measured entities, and measure implementation vendors because they help support measure development, selection, and implementation. This report outlines a framework for integrating AI into quality measurement while maintaining scientific rigor, fairness, and stakeholder trust. The recommendations also offer a foundation for the development of future governance as AI methods evolve and may inform broader applications beyond accountability programs, such as quality improvement and clinical decision support.

Table of Contents

Executive Summary

Quality measures are critical tools that can be used to assess healthcare quality; inform quality improvement; obtain information directly from patients about their outcomes and experience of care; and incentivize high quality care in accountability programs by comparing provider and plan performance; and, if results are publicly reported, directly inform consumer choice. It is important that quality measures used in these "high stakes" accountability applications produce fair, accurate, and consistent results, and processes are in place to support streamlined, high fidelity data collection because providers, patients, payers, and regulators must be confident that measures that are publicly reported and/or used for payment decisions fairly and accurately reflect the aspect of quality they intend to measure.

Artificial intelligence (AI) methods have the potential to accelerate the field of quality measurement by reducing measurement burden, improving reliability and validity of measure scores, accessing and interpreting a broader range of data, and measuring important topics that have not been possible to measure in the past. The current reliance on structured data for measurement imposes burden on clinicians and health systems, due to the need to enter required data into structured fields and navigate cumbersome processes for data extraction and reporting. AI has the potential to unlock the use of high-value clinical data for quality measurement by efficiently accessing and interpreting unstructured data, which enables more comprehensive and patient-centered measures, while lowering burden for providers of care. However, AI-enabled measures may lack transparency into data sources and AI methods and produce biased or inaccurate measure scores. Variation in how providers implement measures could distort results and further erode the trust clinicians, patients, payers, and regulators have in measures used in accountability programs. Federal agencies, collaboratives, and private organizations have released a range of governance and guidance documents on the use of trustworthy AI in healthcare which include themes applicable to quality measurement, but none specifically addresses the use of AI in this context. More guidance is needed to address the novel challenges associated with AI-enabled quality measures and accelerate their use in the context of accountability programs.

To guide and support the use of AI-enabled quality measures, National Quality Forum (NQF), funded by the Gordon and Betty Moore Foundation, convened a multistakeholder technical expert panel (TEP) to drive insights and consensus on guidance that sets standards for developing, selecting, and implementing measures using AI methods in accountability programs. Drawing on the learnings and approaches from broad healthcare AI governance documents and frameworks, the TEP developed six strategies to advance the use of trustworthy AI in quality measurement.

These strategies informed more specific recommendations primarily intended for program owners (i.e., organizations responsible for administering national, regional, state, or local, public- or private-sector accountability programs). Additional key actors have critical roles to play in successful application of this guidance. These include measure developers (i.e., individuals and organizations that develop and test measures), measured entities (i.e., individuals and organizations that collect, report, and are evaluated on quality measure results), and measure implementation vendors (i.e., organizations that assist with measure implementation).

To help individuals and organizations effectively execute recommendations outlined in the strategies, the TEP created an implementation roadmap that details the responsibilities and actions of each key actor. While program owners are positioned to be the primary drivers of this work and can leverage their influence to execute the recommendations, all four key actors play important roles in the successful application of recommendations across the measure lifecycle. The implementation roadmap sets forth recommended actions for five steps in the measure lifecycle for AI-enabled quality measures identified by the TEP: (1) development and testing, (2) selection, (3) preparation for implementation, (4) implementation across entities, and (5) monitoring and maintenance. Table 1 defines each of the five steps. Figure 1 depicts each step in the context of a measure roadmap, listing important activities for key actors and highlighting the feedback loop throughout the lifecycle. To facilitate use of this guidance across the measure ecosystem, NQF and the TEP developed roadmap tables that define the roles of all four key actors during each of these five steps. Finally, the TEP applied a tool to the roadmap tables that is common in project management, called a RACI framework, to define the relative roles (responsible, accountable, consulted, or informed) for actions which require coordination.[1]

Table 1. Description of the Five Steps of the Measure Lifecycle

Key Step	Description
1. Development and Testing	This step involves identifying gaps in measurement and the need for a measure; conceptualizing and sufficiently detailing the measure calculation/specification; and assessing the measure for feasibility, usability, and scientific acceptability. The measure developer considers the advantages and disadvantages of an AI-derived component.* If measure developers choose to include an AI-derived component, they should develop, test, and describe it in the quality measure AI model summary label.
2. Selection	This step begins with a fully developed and tested measure, including its AI-derived component, as a program owner considers using the measure in an accountability program. The program owner reviews measure information (i.e., determines if the measure is important, feasible, reliable, and valid) and the development and testing details, (including the performance of the AI-derived component) to assess the measure's readiness and appropriateness for the program.
3. Preparation for Implementation	This step begins after a program owner selects a measure for use in an accountability program. Activities ensure the measure produces reliable and valid results, and its AI-derived component is generalizable and performs accurately across measured entities before scores are used for accountability decisions (e.g., public reporting, financial incentives). This step will vary by program but may include piloting the measure and its AI-derived component, testing the measure and its AI-derived component, implementing pay for reporting, a measure dry run, and/or voluntary reporting.
4. Implementation Across Entities	This step begins after the program owner determines scores from an AI-enabled measure are accurate, appropriate, and ready to use for accountability decisions. The implementation process involves scaling the measure and its AI-derived component across a large number of measured entities and using results for accountability decisions.

Key Step	Description
5. Monitoring and Maintenance	This step involves monitoring and updating the measure and its AI-derived component as needed. If the measure and/or its AI-derived component change significantly, the measure may need to re-enter the process at one of the preceding steps.

**AI-derived component is a component of a measure identified and/or calculated using AI methods. The terms "AI-derived component" and "component" are used interchangeably throughout the report.*

Figure 1. Five-Step Roadmap for AI-Enabled Quality Measures

The figure below depicts the five-step roadmap for developing and testing, selecting, implementing, and monitoring and maintaining AI-enabled quality measures. Each step highlights key activities grouped by lead actor (i.e., program owner, measure developer, measured entity, measure implementation vendor). However, successful execution of these activities will require collaboration across all key actors. The roadmap begins with (1) development and testing, where measure developers assess feasibility of implementation, rigorously test the measure and its component, design the measure to mitigate for potential gaming, and complete the quality measure AI model summary label. During (2) selection and (3) preparation for implementation, program owners use the quality measure AI model summary label to determine the measure's appropriateness for use in a program. During these steps, program owners also conduct a feasibility assessment, including testing with measured entities, and establish performance standards prior to widespread implementation. During (4) implementation across entities and (5) monitoring and maintenance, measured entities and measure implementation vendors implement, test, and tune the component, as needed; and validate performance against established standards. Entities report these results, and measure developers conduct further maintenance of the component based on feedback.

MEASURE DEVELOPER:

- Evaluates feasibility of the component's implementation
- Rigorously tests the measure and component
- Completes the quality measure AI model summary label
- Designs measure to mitigate potential gaming

PROGRAM OWNER:

- Leverages the quality measure AI model summary label information to determine appropriateness during the selection process
- Performs a feasibility assessment
- Establishes performance standards for the component

MEASURE DEVELOPER:

- Maintains component based on results and feedback from measured entities

PROGRAM OWNER:

- Monitors performance results from measured entities

MEASURED ENTITY AND MEASURE IMPLEMENTATION VENDOR:

- Implement, test, and tune the component as needed
- Validate component performance against performance standards and report results
- Monitor component

SUMMARY OF TEP RECOMMENDATIONS FOR STRATEGIES TO ADVANCE TRUSTWORTHY AI-ENABLED MEASURES IN ACCOUNTABILITY PROGRAMS

1. **Provide precise and transparent information about the AI-derived component, including data used in development and testing:** To support transparency of the AI-derived component, the TEP recommends that measure developers complete a standardized quality measure AI model summary label (see box).[2] The TEP prepared a quality measure AI model summary label template that measure developers should complete. The TEP recommends that program owners should consult the summary label as part of their measure selection process. Measured entities and measure implementation vendors should leverage the summary label to assist with implementation of the component. The TEP additionally recommends that, as part of the development and testing process, developers provide a configuration file for the component and design the measure to avoid the potential for "gaming" (i.e., measured entities manipulating data inputs to obtain optimal performance scores).

   The quality measure AI model summary label includes details such as the rationale for using AI in the measure, the AI-derived component's intended use, descriptions of the data used to develop and test the component, performance results across different populations, and the different risks of the component.[2]

2. **Optimize performance of the AI-derived component through testing and tuning:** The TEP recommends that measure developers follow specific steps when initially testing the performance of an AI-derived component. After measure developers complete sufficient testing of the AI-derived component, the TEP recommends that program owners should establish

standards against which the measured entities should evaluate performance of the component. The TEP further recommends that measured entities tune the component (i.e., adapt the component to local contexts), with assistance from measure implementation vendors, prior to full scale implementation.

3. **Define the capabilities required to implement the AI-derived component (feasibility):** To effectively implement the AI-derived component, the TEP recommends that measure developers identify the resources required to implement the component and document them in the quality measure AI model summary label. The TEP further recommends that program owners perform a feasibility assessment prior to widespread implementation. Measured entities, with assistance from measure implementation vendors, should conduct a feasibility assessment during their initial implementation of the measure. Measured entities should also share their findings with the program owner and measure developer, including the resources and internal and external capabilities needed to implement the component.

4. **Asses regularly for unintended consequences, including bias:** To identify and minimize potential areas of bias and other unintended consequences, the TEP recommends that measure developers and measured entities stratify component performance results by patient characteristics. The TEP also recommends that program owners and measure developers regularly review stratified performance results to assess for bias and other unintended consequences (e.g., impacts on patient safety) in outputs from the component. While the TEP emphasized the importance of stratifying component performance results, members acknowledged it may not currently be feasible in all cases or may be unreliable due to lack of sufficient data or small patient populations.

5. **Prioritize ongoing monitoring and maintenance:** The TEP highlighted the importance of regularly assessing performance of the AI-derived component over time and recommends that measure developers outline a monitoring and maintenance plan, including a cadence for regular updates and feedback collection, in the quality measure AI model summary label. The TEP recommends that program owners monitor and assess component performance on an ongoing basis against established performance standards. Measured entities and vendors should regularly evaluate component performance and provide transparent information about the timing and types of monitoring they conducted for the measure and its component.

6. **Support an ecosystem that enables information sharing and feedback across key actors:** The TEP emphasized the need for a feedback loop process, by which all key actors can communicate and align on information related to component implementation, performance, and maintenance. Given the novelty of this concept, the TEP acknowledged that additional work is needed to facilitate a collaborative approach in which key actors coordinate and work together to establish and maintain this feedback loop.

In addition to their formal strategies and recommendations, the TEP highlighted several emerging topics: sharing code, weights (i.e., a numerical value assigned to a data point or group of data points to reflect its relative importance in producing a model's output)[3] and proprietary details for the AI-derived component; validating the component with a third-party evaluator and/or reference data set; and allowing measured entities to select and apply their own AI method. These topics are crucial to

acknowledge; however, because the use of and considerations around AI are rapidly evolving, the TEP did not reach formal recommendations on these topics. The TEP advised that these issues need further consideration and future recommendations to encourage the trustworthy use of AI in quality measurement while advancing the types of innovative measurement that AI methods allow. Given the novelty of this framework, the TEP recognized this report as an initial step in establishing guidance in the quality measurement field for program owners, measure developers, measured entities, and measure implementation vendors. As the use of AI in healthcare evolves towards more complex methods, the field of quality measurement must assess and update recommendations over time to keep pace with changes.

Introduction

For more than a decade, researchers have investigated the potential uses of artificial intelligence (AI) methods in developing and implementing quality measures.[4–9] It is only in recent years, however, that organizations administering accountability programs (e.g., accreditation, pay-for-performance, public reporting, value-based payment) are considering the regional or national use of quality measures that incorporate AI methods.[10] As further described in the *Background* section, quality measures using AI methods have the potential to reduce measurement burden while allowing significant development in areas that have been previously difficult to measure. While several governance and guidance frameworks for the broad use of AI in healthcare exist, guidance for developing, selecting, and implementing AI-enabled quality measures does not.[11–15] Without this guidance, users of AI-enabled quality measures may be uncertain about the accuracy and trustworthiness of measure results. Cultivating trust is particularly important for measures intended for use in accountability programs. Providers, patients, payers, and regulators must be confident that measures that are publicly reported and/or used for payment decisions fairly and accurately reflect the aspect of quality they are intended to measure.

Project Purpose, Scope, and Approach

INTRODUCTION TO KEY ACTORS AND STEPS IN THE MEASURE LIFECYCLE

To systematically identify guidance and recommendations for developing, selecting, and implementing quality measures that use AI methods, National Quality Forum (NQF) convened a national panel, the Artificial Intelligence in Quality Measures Technical Expert Panel (AI TEP), representing a variety of critical perspectives to drive insights and forge consensus. This work was funded by the Gordon and Betty Moore Foundation. The guidance interprets frameworks focused on AI in healthcare for the use case of quality measurement and identifies the information and actions needed to support the development, selection, and implementation of AI-enabled quality measures.

The aim of this work is to leverage AI's benefits through expanding its use in quality measurement while maintaining scientific validity and trust. TEP members held this aim in view while considering the issues and forming recommendations. The primary audience for this guidance is accountability program owners (Table 2) because they are responsible for implementing quality measurement approaches designed to drive improvements in care while minimizing the burden of collecting and reporting quality measures results. The secondary audiences for this guidance are measure developers, measured

entities, and measure implementation vendors (Table 2), because they are key actors involved in developing, selecting, and implementing quality measures with an AI-derived component (i.e., a component of a measure identified and/or calculated using AI methods for use in accountability programs.* Each of these four key actors has a critically important role in the success of the guidance and recommendations outlined in this report. Each of these four key actors has a critically important role in the success of the guidance and recommendations outlined in this report. Each of these four key actors has a critically important role in the success of the guidance and recommendations outlined in this report.

Table 2. Key Actors Involved in Measure Development and Testing, Selection, Preparation, Implementation, and Monitoring and Maintenance

Key Actor	Description
Program Owners	Organizations (e.g., government agencies, payers, private or non-profit accreditors) responsible for administering national, regional, state, or local, public or private sector accountability programs.
Measure Developers	Individuals and organizations that develop and test measures. If given the responsibility by a measure steward, they may also maintain measures over time and serve as the ongoing point of contact for measure questions.
Measured Entities	Individuals and organizations (e.g., clinicians, clinician groups, health systems, hospitals, health plans) that are evaluated using a specific quality measure. Measured entities are responsible for collecting and reporting quality measure results.
Measure Implementation Vendors	Organizations that assist with measure implementation, including ensuring accurate data collection and providing strategies for performance improvement.

The key actors detailed in Table 2 may fill multiple roles. For example, some measure developers administer accountability programs and some measured entities develop and own measures. Also, the measure developer definition is intended to capture the responsibilities of both measure developers and measure stewards. Measure stewards are another type of key actor involved in measure development, selection, and implementation. They differ from measure developers because stewards own and are responsible for maintaining the measure, although they may assign that responsibility to a measure developer. In some cases, the measure steward is the same individual or organization as the measure developer. In this report, the term "measure developer" is used to indicate both measure developers and stewards. Additionally, program owners will vary by size, type, and degree of influence over a respective program.

To contextualize guidance and recommendations, the TEP identified a five-step measure lifecycle that includes the major phases of developing, selecting, and implementing AI-enabled quality measures: (1) development and testing, (2) selection, (3) preparation for implementation, (4) implementation across

The terms "AI-derived component" and "component" are used interchangeably throughout the report.

entities, and (5) monitoring and maintenance. The five key steps in the measure lifecycle are described in Table 3, while Figure 2 depicts the measure lifecycle, highlighting the feedback loop throughout the lifecycle.

Table 3. Description of the Five Steps of the Measure Lifecycle

Key Step	Description
1. Development and Testing	This step involves identifying gaps in measurement and the need for a measure; conceptualizing and sufficiently detailing the measure calculation/specification; and assessing the measure for feasibility, usability, and scientific acceptability. The measure developer considers the advantages and disadvantages of an AI-derived component. If measure developers choose to include an AI-derived component, they should develop, test, and describe it in the quality measure AI model summary label.
2. Selection	This step begins with a fully developed and tested measure, including its AI-derived component, as a program owner considers using the measure in an accountability program. The program owner reviews measure information (i.e., determines if the measure is important, feasible, reliable, and valid) and the development and testing details (including the performance of the AI-derived component) to assess the measure's readiness and appropriateness for the program.
3. Preparation for Implementation	This step begins after a program owner selects a measure for use in an accountability program. Activities ensure the measure produces reliable and valid results, and its AI-derived component is generalizable and performs accurately across measured entities before scores are used for accountability decisions (e.g., public reporting, financial incentives). This step will vary by program but may include piloting the measure and its AI-derived component, testing the measure and its AI-derived component, implementing pay for reporting, a measure dry run, and/or voluntary reporting.
4. Implementation Across Entities	This step begins after the program owner determines scores from an AI-enabled measure are accurate, appropriate, and ready to use for accountability decisions. The implementation process involves scaling the measure and its AI-derived component across a large number of measured entities and using results for accountability decisions.
5. Monitoring and Maintenance	This step involves monitoring and updating the measure and its AI-derived component as needed. If the measure and/or its AI-derived component change significantly, the measure may need to re-enter the process at one of the preceding steps.

Figure 2. Five-Step Measure Lifecycle for AI-Enabled Quality Measures

METHODOLOGY

This report details the TEP's strategies for advancing trustworthy AI-enabled measures and recommendations for the development, selection, and implementation of AI-enabled quality measures in accountability programs. Because best practices for the development and testing of quality measures designed for use in accountability programs already exist, guidance in this report specifically informs the evaluation and implementation of a measure's AI-derived component.[16–18] The report defines terms to know, briefly provides background on the need for supplemental consensus standards to govern the use of AI in quality measures, and outlines existing AI governance frameworks applicable to this work. The report details six strategies developed by the TEP to advance trustworthy AI-enabled measures and a five-step process to guide the development, selection, and implementation of quality measures using AI methods. Finally, the report outlines the TEP's deliberations on emerging topics.

To produce this guidance, NQF undertook the following steps (described in more detail *in Appendix B*):

1. Convened the multistakeholder TEP (*Appendix A*).
2. Gathered information relevant to the use of AI in quality measures by conducting a review of the literature, existing AI governance documents and frameworks, and consensus-based measure evaluation criteria, and holding several key informant interviews.
3. With the TEP, developed strategies to advance trustworthy AI-enabled measures and recommendations for the development, selection, and implementation of these types of quality measures in accountability programs.
4. Obtain public comment. (Current step)
5. Finalize strategies and recommendations. (Future step)

NQF and the TEP used an iterative process to draft the consensus-based strategies and recommendations (*Appendix B*), incorporating TEP discussions from one in-person meeting and multiple web meetings. In addition, NQF consulted with a five-person advisory group, composed of national leaders with different perspectives that NQF convened to guide the project at key points (*Appendix A*).

The TEP's deliberations were informed by descriptions of measures that already incorporate AI methods; however, the TEP acknowledged that the field of AI generally, and its application to quality measurement specifically, is rapidly evolving, which means that future uses of AI methods in quality measures may expand beyond the use cases the TEP considered in their discussions. To seed conversation, the TEP reviewed a specific measure in detail, the *Diagnostic Delay of Venous Thromboembolism (DOVE) in Primary Care* measure, which uses a rule-based natural language processing (NLP) algorithm to identify venous thromboembolism-related symptoms in clinical notes, while keeping in mind other uses of AI methods in quality measures, including the use of machine learning (ML) and both generative and predictive AI.[19] The TEP chose the DOVE measure to inform its conceptualization of AI-enabled measures because the types of issues that may apply to measures using ML and large language models (LLMs) also apply to measures using rule-based NLP, which currently is more commonly used.

Terms to Know

- **AI-derived component:** A component of a measure identified and/or calculated using AI methods. (Developed by NQF and the TEP)
- **Artificial intelligence (AI):** Refers to the ability of computers to perform tasks that are typically associated with a rational human being—a quality that enables an entity to function appropriately and with foresight in its environment.[20] (Adapted by NQF and the TEP)
- **Generative AI:** Can generate novel text, images, videos, and/or other outputs, typically based on knowledge gained from large datasets.[20] (Adapted by NQF and the TEP)
- **Large language model (LLM):** A subset of generative AI; has the ability to process and/or generate human language.[20] (Adapted by NQF and TEP)
- **Machine learning (ML):** A subtype of AI that involves complex algorithms trained to make classifications and/or predictions about future outcomes.[20] (Adapted by NQF and TEP)
- **Natural language processing (NLP):** A subtype of AI that involves the interpretation and/or generation of text/language.[20] (Adapted by NQF and TEP)
- **Predictive AI:** Uses statistical analysis and machine learning to identify patterns, predict behaviors, and/or forecast future events.[21] (Adapted by NQF and the TEP)
- **Quality measure AI model summary label:** Describes details about an AI-derived component, including the component's intended use, rationale for using AI in the measure, descriptions of the data used to develop and test the component, performance results across different patient populations, and potential limitations and risks.[2] (Adapted by NQF and the TEP)
- **Quality measure**: A standardized tool used to assess the performance of healthcare providers in delivering care that is safe, effective, timely, and patient-centered. These measures help gauge various aspects of healthcare quality and incentivize care improvements. (Developed by NQF and the TEP)

- **Rule-based AI:** Relies on predetermined algorithmic rules to make decisions and/or solve problems. These systems can range from basic pattern matching (e.g., regular expressions) to the use of complex linguistic and ontological methods that guide the AI's actions based on specific conditions.[22] (Adapted by NQF and the TEP)
- **Third-party evaluator:** An independent organization that evaluates outputs and results of an AI-derived component against a third-party data set to ensure the component meets established performance expectations. (Developed by NQF and the TEP)
- **Tuning:** Adapting the component to local contexts through techniques such as hyperparameter optimization; fine-tuning on local data to address distributional shifts; calibration; and post-processing, or prompt-based adaptation. (Developed by NQF and the TEP)

Background

THE NEED FOR SUPPLEMENTAL CONSENSUS STANDARDS TO GOVERN THE USE OF AI IN QUALITY MEASUREMENT

Quality measures can be used for a variety of purposes: to obtain information directly from patients about their outcomes and experience of care; as a feedback mechanism to inform care delivery; and to prioritize and inform quality improvement investments, including reducing disparate outcomes in care across populations, and in accountability programs. When used in accountability programs by healthcare payers and purchasers, measures assess the quality of care providers deliver and/or health plans provide their enrollees. Additionally, measure results may inform which providers health plans include in their networks or which health plans a payer offers and incentivizes its customers to choose. If publicly reported, measure results can directly inform consumer choice. Because performance scores are used to compare provider and plan performance in these "high stakes" accountability applications, it is important that quality measures produce fair, accurate, and consistent results over time and across measured entities that may serve highly variable patient populations.

> National, consensus-based criteria help ensure that quality measures are based on evidence and are associated with gaps or variations in care (important), are consistent across time and measured entities (reliable), accurately represent the evaluated concept (valid), and are based on data and resources available to measured entities without undue burden (feasible). These requirements or criteria confirm that measures and their resulting scores can be used for national or widespread comparisons across entities and drive improvements in care.[16]

To achieve widely accepted, scientifically sound measures, the field applies consensus-based criteria (see box) developed through structured processes that provide transparency, support technical evaluation, and engender trust. NQF as the initial national consensus-based entity, focused on advancing quality through measurement with multistakeholder input and authored initial national endorsement criteria in 2000, and continues to provide supplementary recommendations to address emerging issues in the field to better enable the use and positive impact of quality measurement.[23] These criteria are used in national endorsement processes and inform payers, providers, and purchasers across public and private entities as they consider the potential use of quality measures in accountability programs.[16] These consensus-based criteria create a strong foundation for integrating technologies such as AI into the quality measurement landscape.

THE PROMISE AND THE CHALLENGES OF AI-ENABLED MEASURES

AI holds strong potential to accelerate progress toward quality measures that are low burden to implement, highly reliable and valid, aligned with patient priorities, and capable of providing real-time feedback to drive improvement. AI methods can efficiently access and interpret a broad range of healthcare data required for patient-centered measures and lower the effort required for data acquisition and validation, potentially reducing measurement burden. AI methods may also enable measurement of important topics that were previously difficult to assess due to the unstructured nature of relevant data. For example, the use of AI in quality measurement enhances the ability of quality measures to leverage the full set of data available on patient care (e.g., pulling free-text data from clinical notes and laboratory or radiology reports using NLP). These data sources are currently burdensome and time-consuming to review and extract, often requiring specialized training for human abstractors to ensure accuracy and consistency. AI may also enable capture of data not directly documented in the electronic health record (EHR) but rather transmitted into it by external devices such as ambient sensors, voice recorders, or bedside electrocardiogram (ECG) monitors. For instance, ambient voice recordings of the patient interview enable capture of the patient's history and symptoms in their own voice, complementing clinicians' interpretations. These enriched data can support more comprehensive measures of diagnostic quality. In addition, ML can differentiate between patient subgroups and complex clinical scenarios with greater precision, making quality comparisons across diverse providers more feasible. AI methods can also find data in its native location, gather data from disparate sources, and normalize data.

While AI offers significant benefits for quality measurement, it also introduces new governance challenges in the development, selection, and implementation of quality measures, similar to those encountered with other technologies when they were first used for quality measurement (e.g., EHRs). A central concern is transparency. For example, the methods used to develop AI algorithms, the data used for development and testing, and/or the measure specifications may not be fully transparent. Yet transparency of the development process, development and testing data, and measure calculation logic is a key expectation that providers, consumers, payers, and others have of quality measures used in accountability programs. Currently, panels that review measures against consensus-based criteria have complete visibility into measure specifications and logic, including the patients included and excluded from the measure, how clinical processes or outcomes are measured, and which variables are used in predictive models for risk adjustment.

Complete transparency of the measure's details and its relationship to clinical evidence, as well as the development process, supports reviewers' assessments of face validity (i.e., the extent to which a measure appears to cover the concept it intends to assess).[24] Transparency of the measure also allows program owners to evaluate the potential applicability of a measure for an accountability program and measured entities to understand the measure and implement it correctly. Transparency also supports the auditability of the measure. However, there are legitimate reasons why full transparency may not be possible for an AI-derived component specifically. A measure developer may use an algorithm that is proprietary, developed and owned by a third party, and functions as a "black box," with the underlying AI code not accessible to users. For example, an LLM that is not open source may limit the developer's ability to disclose details about the algorithm's design or the data used to train it. Similarly, when a measure uses an ML-based risk adjustment model, the underlying logic may not be fully accessible if the model does not make the evaluated variables and weights (i.e., a numerical value assigned to a data

point or group of data points to reflect its relative importance in producing a model's output)[3] transparent. These challenges highlight the need for further guidance that sets realistic expectations for transparency while supporting program owners in evaluating measures against consensus-based criteria.

In addition to transparency challenges, another concern is the potential for AI-enabled measures to perpetuate disparities in care. These measures may unintentionally set lower expectations for care delivery and outcomes for certain patient groups, particularly if they rely on historical performance patterns that reflect existing inequities.[25] Because humans influence the design of AI models and the data used to train them, there is a risk of both implicit and explicit bias, especially against historically marginalized groups.[26] This risk is heightened when training data are incomplete or underrepresent groups that experience fragmented care due to limited access to clinical services. Differences in medical treatment and diagnosis can also lead to biased or poorly representative data sets.[27]

Adequate recognition and management of these risks of bias is necessary to prevent harm. However, evaluating the generalizability of AI-enabled measures can be difficult. When ML or LLMs are used to develop an algorithm or to pull data from patient records, the training inputs may not reflect aspects of the provider or patient population for which the measure is intended. In many cases, the attributes of the development and testing data sets are not available, making it difficult to assess whether the measure is appropriate for its intended population. To address this, there is a need for consensus on what testing and descriptive data should be available for program owner review. To minimize the potential introduction of bias, it will be important for measure developers and program owners to periodically evaluate the risk of biases for a component's outputs and have a plan for addressing biases once identified.

Assuring that implementation of quality measures across entities produces consistent and valid results is also critical to ensure validity, reliability, interpretability, and fair comparison of performance scores. Yet this goal may be more challenging for measures using AI methods. Measure developers traditionally provide precise and complete specifications for measure calculation and conduct reliability testing to demonstrate that the measure can produce consistent and comparable results. While AI-enabled quality measures may face challenges in generalizability, this potential limitation should not discourage their use. Instead, it underscores the importance of designing and implementing these measures to be effective across diverse settings and patient populations.

Measures with AI-derived components developed and trained in settings with specific system-level variables (e.g., clinical documentation practices, type/availability of clinical notes) and patient-level variables (e.g., insurance type, comorbidities, age, race/ethnicity) may not automatically translate well to other contexts. Without careful consideration, this can limit the ability to evaluate, interpret, and compare performance scores.[28] Furthermore, component adjustments or tuning at the local level could compromise the comparability of benchmarks used for payment or public reporting. Processes are needed to verify that AI-enabled measures are implemented with fidelity and produce comparable results across measured entities and patient populations.

There are additional concerns related to the feasibility of implementing and scaling an AI-enabled quality measure. Feasibility is a crucial characteristic for quality measures used in accountability programs. It is important for program owners to assess feasibility prior to widespread implementation

of a quality measure to understand the internal and external capabilities measured entities may need to report the measure and help support consistent implementation of AI methods across entities. Considering the novelty and added complexity of measures that use AI methods, measured entities may encounter challenges related to limited resources (e.g., financial resources, computational power, specialized infrastructure) and expertise (e.g., skilled professionals trained in implementation and use of AI models) needed to appropriately implement an AI-derived component.

Such constraints are likely to be more pronounced and have a more significant impact on smaller or under-resourced measured entities (e.g., rural systems) that may not have the same access to technical capabilities and expertise as larger measured entities (e.g., urban academic health centers). As such, the landscape of AI in quality measurement is likely to remain uneven, not only across measured entities, but also across different types of quality measures depending on how interdisciplinary or discipline-specific they are. For example, the use of AI in radiology may be more common than other specialties. This potential variability in which types of measures get to benefit from AI methods could create challenges in measuring quality across disciplines.

In summary, when a program owner considers a measure for potential use in an accountability program, their review should assess whether the measure is aligned with the program's aims and scope; is appropriately defined and tested for the relevant patient population, care settings, and measured entities; and demonstrates its usability and potential to drive improvements in care. These foundational considerations remain essential even when a measure leverages AI methods. However, more specific guidance is needed to support review and assessment of the AI methods used in the measure to address the concerns highlighted above.

EXISTING AI GOVERNANCE FRAMEWORKS AND THEIR IMPLICATIONS FOR QUALITY MEASURES THAT USE AI METHODS

With the recent proliferation of AI development and implementation in healthcare and other fields, there is increased attention on governing and guiding the responsible use of trustworthy AI.[29,30] Many federal agencies, collaboratives, and private organizations have released frameworks describing how to develop and implement trustworthy AI algorithms and models.[12–15,25,31,32] The Light Collective, a patient advocacy organization, has also published a framework for the use of trustworthy AI in healthcare focusing on patients' rights.[33]

Because these frameworks focus on the use of AI in healthcare generally, the quality measurement field needs new interpretative guidance to apply specifically to AI-enabled quality measures. While the underlying expectations and requirements for measures will not change with the introduction of AI methods, measure developers who create these measures will need to continue to provide specific details about the AI-derived component beyond what they typically provide, including a detailed description of the component, data used for development and testing, and results from testing the performance of the component. Developers will also need to demonstrate how the component can be feasibly implemented across various care settings and measured entities included within applicable accountability programs.

NQF's approach to this project was to apply principles emerging in the national dialogue about the use of AI in healthcare to the use case of quality measurement and leverage existing frameworks to inform

and develop guidance and recommendations for AI-enabled quality measures. To inform NQF's discussions with the TEP, NQF reviewed and shared learnings with the TEP about several national governance documents and frameworks, including the following:

- National Institute of Standards and Technology (NIST) Risk Management Framework[31]
- The Assistant Secretary for Technology Policy (ASTP) (formerly the Office of the National Coordinator for Health IT [ONC]), Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) Final Rule (ONC HTI-1 Final Rule)[12]
- U.S. Food and Drug Administration (FDA) Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)[13]
- FDA, Health Canada, and the United Kingdom's Medicines and Healthcare Products Regulatory (MHRA) Agency Good Machine Learning Practice for Medical Device Development: Guiding Principles[34]
- FDA Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations Draft Guidance for Industry and FDA Staff[35]
- National Academy of Medicine (NAM) An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action An AI Code of Conduct Principles and Commitments Discussion Draft[36]
- Consumer Technology Association (CTA) Artificial Intelligence in Health Care: Practices for Identifying and Managing Bias[37]
- CTA The Use of Artificial Intelligence in Health Care: Trustworthiness[38]
- Coalition for Health AI (CHAI) Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare[32]
- CHAI Responsible AI Guide[54]

In addition, NAM presented its draft 2024 AI Code of Conduct for Health, Health Care, and Biomedical Science to the TEP early in the project.[11] This presentation highlighted key learnings that emerged from a landscape review supporting the Code of Conduct. A systematic literature review of 56 documents on "socially responsible AI" found that many of these documents included fairness and transparency as key themes.[39] The literature review included scientific literature published between 2018-2023 that focused on responsible AI principles; guidance developed by medical specialty societies for physicians using AI; and frameworks, policies, and guidance issued by the federal government through May 2023.

Appendix C provides a summary of the key learnings from each framework and how the framework informed NQF's work with the TEP.

Other guidance documents and emerging activities also informed the TEP's discussions about the trustworthy use of AI in quality measures. For example, the TEP discussed the use of model summaries to provide transparency about the AI-derived component. Several organizations have started to produce templates for brief summaries of AI models.[35,40,41] These summaries offer details about an AI model, supporting assessments of the model's intended use, performance across different patient populations, and risks.[18] The FDA included an example summary template for developers of AI-enabled device software in their Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and

Marketing Recommendations Draft Guidance for Industry and FDA Staff.[35] The CHAI summary template aligns with the ONC HTI-1 Final Rule for decision support interventions.[12,40] Additionally, the MINimum Information for Medical AI Reporting (MINIMAR) proposal acknowledges the need for reporting standards on AI in healthcare and outlines minimum information needed to understand AI predictions, target populations, biases, and generalizability.[42]

TEP Recommendations for Strategies to Advance Trustworthy AI-Enabled Measures in Accountability Programs

Drawing on themes in existing AI governance and frameworks, the TEP first agreed on six strategies to advance the use of AI-enabled quality measures in accountability programs. These strategies informed their recommendations for developing, selecting, and implementing quality measures that incorporate AI methods. To make the recommendations as actionable as possible, the TEP organized strategies by the roles and responsibilities of the four key actors—measure developers, program owners, measured entities, and measure implementation vendors—defined previously (Table 2)**.** The following sections include a bulleted list of responsibilities for each of these individuals and organizations. Similarly, the TEP developed a *Roadmap for Implementing TEP Recommendations* which enumerates recommended actions by each key actor for each step in the lifecycle.

Unless a fact or recommendation is explicitly attributed to a specific source, information in the rest of the report comes from the TEP and was synthesized by NQF.

STRATEGY 1: PROVIDE PRECISE AND TRANSPARENT INFORMATION ABOUT THE AI-DERIVED COMPONENT, INCLUDING DATA USED IN DEVELOPMENT AND TESTING

TEP Recommendations

- Measure developer:
  - Completes the quality measure AI model summary label for the AI-derived component
  - Provides a configuration file for the AI-derived component and other information that aids implementation
  - Designs the measure and its AI-derived components to avoid the potential for "gaming" (i.e., measured entities manipulating data inputs to obtain optimal performance scores) and describes their approach for the program owner
- Program owner:
  - Consults the quality measure AI model summary label provided by the measure developer as part of their measure selection and implementation process
- Measured entity with assistance from measure implementation vendor:
  - Consults the quality measure AI model summary label as part of its implementation process

Standardized Quality Measure AI Model Summary Label

Program owners should consult the quality measure AI model summary label completed by measure developers as part of their measure selection and implementation process. This quality measure AI model summary label will facilitate communication between the measure developer and the program owner, measured entities, and measure implementation vendors. It also provides important information

for a program owner to consider about the appropriateness of the measure and its AI-derived component for use in a program. Measured entities and measure implementation vendors can use the label to understand the requirements to implement the AI-derived component (e.g., the data/inputs needed to implement the component, the generalizability of the component to their patient population, and the risks of implementation).

Measure developers may also have additional information that will assist measured entities with implementing the AI-derived component and assessing the performance of the component once it is in use. Examples include providing synthetic notes to sites implementing an LLM or instructions on how to use an NLP-based component accompanied by a sample of de-identified notes. While transparency is important for AI-enabled measures being considered for widespread implementation in accountability programs, the TEP acknowledged developers may have difficulty being completely transparent about proprietary AI models developed by third parties. This is discussed in more detail in *Emerging Topics*.

Measure developers should complete the quality measure AI model summary label, using the template developed by the TEP. This template leverages the summary template outlined in the FDA's Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations Draft Guidance for Industry and FDA Staff and is informed by the CHAI summary template.[35,40] The TEP noted that these existing model summaries may be missing important details (e.g., information about data pre-processing) and changes to these examples should be monitored for potential revisions to this template. The TEP also acknowledged that measure developers may need additional guidance about whether to complete quality measure AI model summary label fields for the AI-derived component as a whole or variables within the component. Because of the number of fields in the quality measure AI model summary label, the TEP identified fields they considered to be high priority, which are indicated by an asterisk. Additionally, a TEP member provided a completed example of the quality measure AI model summary label, using an NLP use case. *Appendix D* includes this example.

Quality Measure AI Model Summary Label Template for AI-Derived Component in a Quality Measure

AI-Derived Component Information

- Name of the AI-derived component
- Name of the developer of the component (may or may not be the measure developer)
- Version of the component used in the measure (i.e., model/software release version)
- Date when the component was created (or last updated)

Description

- Intended users (e.g., healthcare providers, health plans, caregivers, patients)*
- Intended use: The general purpose of the component or its function. This includes descriptions of how the component is used in the quality measure, the target patient population for which the component is intended, and the intended care setting(s) in which the component is used (e.g., hospital, ambulatory care)*
- Instructions for use: Directions and recommendations for optimal use of the component in the measure by the measured entity*

- Rationale: The rationale for using the component in the quality measure, including a description of the clinical or quality concept that it attempts to capture, why AI should be used to capture the concept rather than other methods (e.g., administrative data, electronic health record [EHR] data), and how the resulting definitions, associated codes and terms, variables, and other inputs represent the clinical concept*
- Type of algorithm/model, including whether the component is predictive or generative, and a description of how it interacts with other systems (e.g., EHRs, integrated platforms, patient-generated information)*
- Inputs: A description of the data source(s) used as inputs by the component, including the source of data that are necessary as input into the component and the types of data used (e.g., EHR, imaging)*
- Outputs: A description of the outputs of the component, including the type and value, and whether the output is a prediction, classification, evaluation, analysis, or another form**

Development and Testing

- Characterization of data used to develop and test the component (these data sets should be separate)*:
    - Data sources (e.g., health system data, public or proprietary databases) including details on any devices used to collect data*
    - Data types used (e.g., structured numerical data, structured categorical data, unstructured text, images, time-series data, or combinations of data types)*
    - Pre-processing applied to data before developing the component*
    - Relevant details including*:
        - Unit of analysis*
        - Number of patients/records/data points*
        - How the developer sampled the data, if applicable*
        - A description of the data sources that were available in the data set but not included and why the developer did not include them*
        - Characteristics of patients included in the data set*
        - Characteristics of patients excluded from the data set*
    - A description of subpopulation characteristics (e.g., the percentage of subgroups captured by the component) and an assessment of whether the data can be considered representative of the overall intended population*
    - Characteristics of healthcare entities included in the data set*
    - Characteristics of healthcare entities excluded from the data set*
    - A description of the process for developing the component*
- Description of how missing data and/or a limited data set may impact performance of the component*
- Limitations of the data sets used for development and testing, including if the developer needed to normalize or translate the data*

Performance

- A description of the process used for testing the performance of the AI-derived component and a description of the types of tests used*

- A summary of the performance results*
- Stratification of the testing results by patient characteristics*
- Links to published evidence describing development and/or testing of the AI-derived component*

Risk Management

- Potential risks associated with the component, the data, and the outputs (e.g., bias risks, information gaps)
- Interactions, deployment, and updates. When appropriate, provide the:
    - Resources required to implement the component, including computational resources, IT infrastructure, staffing expertise and numbers, and whether there is a cost to license the component
    - Details regarding how the component is deployed and updated, including:
        - How to conduct local site-specific acceptance testing or validation
        - Ongoing performance monitoring and maintenance
        - Transparent reporting of successes and failures
        - Change management strategies
        - Proactive approaches to address vulnerabilities
    - Communication to parties of as-needed information
    - Software quality (specify standards and regulatory compliance issues, intellectual property issues, risk management and safeguards used)
- Known risks, biases, or failure modes
- Bias mitigation approaches used during development and testing of the component
- Known circumstances where the input for the component will not align with the data used in development and validation
- Ethical or clinical implications that may arise from component misclassification

Configuration File

Measure developers should provide a configuration file for the AI-derived component. Because the quality measure AI model summary label on its own will not provide enough information for a measured entity to implement an AI-derived component, the measure developer should provide a configuration file. A configuration file is a machine-readable document (e.g., .yaml, .json, or .ini) that captures, in one place, every setting needed to reproduce a given model run, including data sources, preprocessing steps, model-architecture choices, training hyper-parameters, hardware/environment settings, and evaluation metrics. Storing these parameters outside the code base enables exact reproducibility across time and environments, promotes transparency and auditability (especially when version-controlled), and allows non-developers to inspect or adjust parameters without editing code.

Design the Measure to Avoid Potential for Gaming

Measure developers should design the measure to avoid the potential for gaming. As with other types of quality measures, there is a risk for gaming. The use of AI introduces distinct risk that end users can influence outputs with novel techniques such as “model manipulation” (i.e., the intentional influence of an AI system’s behavior or outputs in a way that deviates from its intended function).[43] These vulnerabilities make it crucial for measure developers to protect against gaming, to the extent possible, particularly if a measure is used in an accountability program. The TEP recommends that measure developers review studies and/or evidence of previous gaming scenarios in measurement and describe

their approaches to minimizing this potential, which may include outlining potential scenarios, metrics used to assess for gaming, how developers designed the measure to counteract these risks, potential measure vulnerabilities to gaming, and how the program owner can monitor the measure results for gaming. **Developers should clearly document and share this information with the program owner.**

STRATEGY 2: OPTIMIZE PERFORMANCE OF THE AI-DERIVED COMPONENT THROUGH TESTING AND TUNING

TEP Recommendations

- Measure developer:
    - Tests the measure according to existing consensus-based criteria and tests the AI-derived component with a different data set than the one used for development
    - Conducts testing of the AI-derived component performance according to current industry standards
    - Provides performance metrics for the AI-derived component (included in quality measure AI model summary label)
    - Provides guidance to measured entities and measure implementation vendors about testing they should conduct locally when implementing the measure
    - Provides guidance to program owners on appropriate performance standards against which measured entities compare component performance results
- Program owner:
    - Establishes performance standards for the AI-derived component, including expected accuracy and precision
    - Ensures measured entities and measure implementation vendors compare performance of the component against performance standards and reviews performance results
    - Sets requirements for measured entities to document steps taken to locally tune the component (i.e., adapt the component to local contexts) and provide performance results of the component before and after tuning
- Measured entity with assistance from measure implementation vendor:
    - Tests the performance of the AI-derived component when they implement the measure
    - Compares results of the component against the established performance standards and tunes the component, as needed, to meet performance standards and retesting performance after tuning
    - Reports performance results and steps taken to tune the component to the program owner and measure developer
- Measure implementation vendor
    - Supports measured entities to validate implementation

A "Cultural Change" in Measure Testing and Implementation: Greater Shared Responsibility

Current quality measures are typically tested by the measure developer in accordance with existing consensus-based criteria to ensure they are reliable and valid. Measure developers must show the data used are reliable and as extracted have comparable meaning across sites so that measure scores will be comparable. The level of testing varies by data type (e.g., audited claims are often assumed valid while electronic clinical quality measures [eCQMs] using EHR data are tested in at least two sites with different

EHRs). Measures do not typically undergo additional testing once they are implemented by measured entities.

For AI-enabled quality measures, the TEP advised that measured entities, along with measure implementation vendors, will need to provide additional oversight of AI-derived components, due to the novelty and rapid evolution of these AI technologies. These components will likely need to be tuned by measured entities to account for variations in local data and workflows. In addition to local tuning, the TEP recommends that measured entities and measure implementation vendors validate and share the component's results with program owners and measure developers, a step that is not currently required for traditional quality measures. The TEP acknowledged that these additional requirements for testing and tuning the AI-derived component will be a perceived "cultural change" in quality measurement and recognized the tradeoff between imposing burden on key actors involved in the process and gaining greater transparency of AI-derived component results.

As the AI landscape progresses, it may be onerous for key actors, particularly measured entities, to fully implement these recommendations. As a result, the TEP advised that the recommendations should be considered best practices. However, because of the importance of and need for local testing and tuning, the TEP encouraged the quality measurement field to find avenues to support less-resourced measured entities so that they can also benefit from the advantages of AI-enabled quality measures even if some entities are not able to engage in the same level of testing and tuning as others.

Initial Testing of the AI-Derived Component by the Measure Developer

Measure developers should adhere to existing measure testing requirements as outlined by consensus-based measure evaluation criteria, while also conducting specific testing of the AI-derived component. To guarantee performance scores that fairly and accurately reflect quality differences across diverse measured entities, measure developers currently demonstrate that the measure is well-defined and precisely specified to enable consistent implementation within and across groups. Reliability and validity testing should also demonstrate that the measure's data elements are repeatable, the performance score is precise, and the measure assesses the quality concept it intends to assess.

TEP members noted that measure developers may choose to test the AI-derived component for reliability (e.g., test-retest, inter-rater consistency) and validity (e.g., expert review, correlation with known indicators) to demonstrate the component's clinical and operational relevance. However, to support repeatability and clarify performance results, TEP members recommend that measure developers provide the level of granularity at which they conducted testing on the AI-derived component. In this context, granularity refers to how the component's performance is evaluated and reported, including whether performance metrics are reported for each individual output or for a composite outcome. For example, a component developed to detect multiple types of complications might be tested for its ability to identify each complication separately or, alternatively, the measure developer might report a single, aggregated score indicating the component's ability to detect any complication.

The TEP noted a specific challenge for developing, testing, and then later monitoring AI-derived components that are assessing rare events: random samples may not contain enough events to allow

appropriate modelling or testing. TEP members cautioned that the results of testing related to rare events may be hidden by accuracy or averaged metrics and suggested developers may need to use non-random, high likelihood samples to test all necessary events captured by the AI-derived component.

Types of Testing for the Measure Developer to Conduct

Measure developers should test the AI-derived component according to current industry standards and provide the results of the testing and the performance metrics used to test the component in the quality measure AI model summary label. The TEP recommends performance metrics that measure developers could leverage for binary and continuous variables using current industry standards, although members recognized these tests could change with advancements in AI. Table 4 provides examples with definitions of performance testing measure developers could conduct to assess the AI-derived component. These tests are not currently required and should be considered best practices for testing an AI-derived component. However, as the use of AI-enabled measures in accountability programs progresses, these recommendations may help inform current consensus-based entity testing requirements. As industry standards for the performance metrics that developers should use test AI methods evolve, the TEP emphasized the importance of reporting quantitative metrics of evaluation against current industry standards to confirm that the AI-derived component remains aligned with the rapidly changing AI landscape.

Table 4. Types of AI-Derived Component Performance Testing for Binary and Continuous Variables

Relevant Variable	Performance Tests

Binary Variables	• **Area Under Receiver Operating Characteristic Curve (AUROC):** Measures how well a model can differentiate between positive and negative classes. It is calculated by area under the ROC curve, which plots true positive rate against the false positive rate (1 – specificity) across various thresholds. Higher AUROC indicates better discriminative performance.[44]
	• **Area Under Precision-Recall Curve (AUPRC):** Measures how well a model can classify positive classes, especially when data are imbalanced and positive classes are rare. It is calculated by area under the PR curve, which plots precision against recall. Higher AUPRC indicates better performance in identifying positive classes.[45]
	• **Positive Predictive Value (PPV):** Also known as precision, PPV is the proportion of true positive predictions among all positive predictions and indicates the accuracy of positive predictions made by a model.[46] PPV is calculated as:
	  ○ PPV = True Positives / True Positives + False Positives
	• **Negative Predictive Value (NPV):** The proportion of true negative predictions among all negative predictions and indicates the accuracy of negative predictions made by a model.[46] NPV is calculated as:
	  ○ NPV = True Negatives / True Negatives + False Negatives
	• **Sensitivity:** Also known as recall or true positive rate, sensitivity is the proportion of true positives that are correctly identified by the model and indicates the model's ability to detect true positives[46] Sensitivity is calculated as:
	  ○ Sensitivity = True Positives / True Positives + False Negatives
	• **Specificity:** Also known as true negative rate, specificity is the proportion of actual negatives that are correctly identified by the model and indicates the model's ability to detect true negatives.[46] Specificity is calculated as:
	  ○ Specificity = True Negatives / True Negatives + False Positives
	• **F-1 Score:** Measures accuracy of model classification, especially when data are imbalanced. It is calculated as the harmonic mean of precision and recall.[47] F-1 is calculated as:
	  ○ F-1 = 2 X (Precision X Recall / Precision + Recall)

Relevant Variable	Performance Tests
Continuous Variables	• **Coefficient of Determination (R^2):** The proportion of variance in the dependent variable that is predicted by the independent variable(s). It ranges from 0 to 1, where 0 indicates the model explains no variability and 1 indicates the model explains all variability.[48] • **Mean Squared Error (MSE):** The average of the squared differences between actual and predicted values. MSE penalizes larger errors more than smaller ones due to squaring. Lower MSE indicates that the model's predictions are closer to the actual values, signifying higher accuracy.[48] • **Mean Absolute Error (MAE):** The average of the absolute differences between actual and predicted values. MAE provides a more straightforward interpretation of prediction error in the same unit as the target variable. Lower MAE indicates that the model's predictions are closer to the actual values, signifying higher accuracy.[48] • **Root Mean Square Error (RMSE):** The square root of the MSE. It indicates the average difference between a model's predicted values and actual values.[48] • **Pearson's R:** Measures the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.[49]
Binary and Continuous Variables	• **Calibration Curve Assessments:** Graphical representation that illustrates the relationship between a model's predicted probabilities and actual observed outcomes and assesses the quality of a model's probability predictions. Several of the tests contained in the table above are calibration curve assessments.[50]

The TEP had additional detailed discussion about the role of human chart abstractors in testing an AI-derived component. A TEP member noted that a measure developer may want to first conduct inter-rater reliability testing of a data element within a measure using human chart abstractors before moving to comparisons of human chart abstractors and AI abstraction because it will help the measure developer establish that the data element is well-defined. Otherwise, it may be difficult to determine if poor inter-rater reliability between the human and AI abstractors indicates an issue with the measure specification itself or the AI-derived component. Another TEP member cautioned that the "gold standard" against which the measure developer compares the AI-derived component does not always have to be human abstraction because there are instances where AI may be more accurate than the human abstractor. The TEP suggested an "adjudication step" during testing, in which the measure developer compares differences between the results from a human versus an AI chart abstractor to identify which is more accurate and should be considered the gold standard.

Measure developers should develop the component using separate measured entities or data sets than those used for testing, which aligns with a best practice for ML.[15] The tested systems or data sets should be sufficiently different with regard to patient population, setting, region, and EHR systems. Testing with a different data set or site than the one used for development protects against

idiosyncrasies specific to individual organizations and sites and allows for higher confidence that the AI-derived component will perform well when implemented across a larger number of entities. The TEP recognized there may be limitations on the amount of testing measure developers are able to conduct due to cost and other resource constraints, yet TEP members advised that measure developers should complete testing using more than one data set or entity. The TEP did not reach agreement on a specific number of data sets or entities. However, members emphasized that implementing AI models can itself be costly, and AI-derived components may introduce bias to measure results, which highlights the need for rigorous testing. Variations in clinical practice, documentation, and data, and the limited implementation experience with AI of many measured entities may make these results inaccurate or inconsistent, underscoring the importance of testing across multiple measured entities and/or data sets and being transparent about the measured entities and data sets used to test the component.

Performance Testing of the AI-Derived Component by Measured Entities

In addition to the initial measure developer testing of the AI-derived component, **measured entities should test the AI-derived component when they implement the measure with support from measure implementation vendors.** This testing verifies that the component performs as intended when it is implemented. Local validation is necessary when measured entities implement an AI-derived component because entities may not be able to implement the component exactly as defined by the measure developer. Differences in clinical practice, workflows, and documentation; data capture and storage; clinician and patient vocabulary; and other idiosyncrasies impacting the data inputs for an AI-derived component necessitate validation of a component's outputs by each measured entity.

Measure implementation vendors will likely play a large role in assisting and guiding measured entities through implementing the component, including validating outputs. For instance, the TEP recommends that **measure implementation vendors may work on behalf of measured entities to guide and validate implementation.** This support from vendors would streamline the validation process and help confirm the AI-derived component functions properly within local contexts and environments. To support evaluation of the component's performance and encourage accurate implementation, **measure developers should provide guidance to measured entities and measure implementation vendors about which testing to conduct during local validation.**

To ensure that measured entities implement the component accurately and consistently with the measure's intent, based on guidance from measure developers, **program owners should establish performance standards against which measured entities, with support from measure implementation vendors, compare component results.** These performance standards set thresholds or ranges against which measured entities compare their own performance results of the AI-derived component.

The TEP identified complexities with implementing AI-derived components which need attention, specifically LLMs. Members noted how a component based on an LLM may deliver variations in outputs even with the same prompt and cautioned that program owners will need to identify whether measured entities should report results using the worst, best, or average of these different outputs. If the program owner does not specify requirements related to this, measured entities may choose the outputs that provide them with the best performance. The TEP also considered whether program owners would require all measured entities to achieve the same performance result for any type of AI-derived

component (e.g., an accuracy of 90 percent) or whether owners would be willing to accept variation in performance if it was over a certain threshold.

The TEP also discussed how sensitivity analyses may help to identify the impact of differences in data inputs on the performance of AI-derived components. For example, members mentioned how academic medical centers may have different clinical documentation practices than community-based settings, and this may affect how the component performs in these different settings. Program owners may be interested in conducting a sensitivity analysis to better understand the impact of the inputs on the output's performance and also the component's generalizability. The measure developer could inform the sensitivity analysis because of their understanding and experience with the component.

Tuning of the AI-Derived Component by Measured Entities and Tuning Limits

Program owners should set requirements for measured entities to document steps taken to locally tune the AI-derived component and provide performance results of the component before and after tuning. To meet the performance standards established by the program owner, measured entities may need to perform tuning of the AI-derived component. Tuning, in this context, is defined broadly as adapting the component to local contexts through techniques such as hyperparameter optimization, fine-tuning on local data to address distributional shifts, calibration and post-processing, or prompt-based adaptation. Some TEP members suggested that tuning should be performed at the highest possible level of aggregation (e.g., at the health plan or system level, rather than the individual clinician or hospital level). Additionally, if tuning occurs, members noted that any statistics (e.g., validation) previously provided for the component will no longer be applicable. **Measured entities, with support from measure implementation vendors, should compare performance of the tuned component against the established performance standards and report results to the program owner.**

As the use of AI-enabled measures progresses, it becomes increasingly important to understand how these components are implemented across measured entities. Many measures currently have explicit requirements and specifications that measured entities must meet exactly, following the same specifications and using the same codes as all other measured entities. However, even with current accountability measures, there are differences in implementation that are not transparent. For example, even though eCQMs use standardized code sets, the data may be extracted from different data fields or sources. In contrast, the data for medical record review measures rely on chart abstractors who demonstrate measurable differences in the interpretation of measure specifications. Measures with AI-derived components may have more apparent deviations in implementation because measured entities may tune the components, increasing the need for transparency around component performance for each measured entity.

TEP members also noted that local tuning of the AI-derived component may make it difficult to compare performance across sites using measure scores. However, members also agreed that local tuning must occur to guarantee that the component performs as intended. TEP members acknowledged difficulties in enumerating a recommendation to minimize this challenge due to the unknown impact of the tuning process across sites, and the novelty of measures incorporating AI-derived components.

STRATEGY 3: DEFINE THE CAPABILITIES REQUIRED TO IMPLEMENT THE AI-DERIVED COMPONENT (FEASIBILITY)

TEP Recommendations

- Measure developer:
  - Identifies and describes the resources required to implement and validate the AI-derived component (included in quality measure AI model summary label)
- Program owner:
  - Performs a feasibility assessment, including testing with measured entities, prior to widespread implementation of the measure with an AI-derived component
- Measured entity:
  - Conducts a feasibility assessment, including resources and internal and external capabilities needed to implement the component, and shares these findings with the program owner and measure developer
- Measure implementation vendor:
  - Supports measured entities in conducting a feasibility assessment, determining the most efficient way to implement the component, and shares these findings with the program owner and measure developer

Identifying Resources Required to Implement the AI-Derived Component

Measure developers should identify and describe the resources required to implement and validate the AI-derived component, including information on the number and expertise of staff required to implement and monitor the component's performance, additional information related to technology infrastructure (e.g., computing systems, data storage) measured entities need at the time of evaluation and throughout implementation of the component, and whether there is a cost to license the component or fees associated with using the component. Because the use of AI is still relatively new in healthcare generally and particularly in quality measurement, many entities collecting data for and reporting on measures leveraging AI-derived components are less likely to have expertise in implementing AI, particularly as the measurement field moves beyond NLP to more advanced AI models, such as LLMs. To support standardized and effective implementation of the AI-derived component, it is critical for measure developers to provide clear, transparent information on what measure entities will need. This detailed information will help program owners, measured entities, and measure implementation vendors to assess readiness, plan, and allocate resources effectively. Such guidance will also reduce implementation burden and support the consistent use and application of AI-derived components assessing the same clinical concept.

Performing a Feasibility Assessment

Program owners should perform a feasibility assessment of the AI-derived component, including testing with measured entities prior to widespread implementation. The initial feasibility information may come from the measure developer's quality measure AI model summary label and then may be supplemented by feedback from measured entities as they implement the AI-derived component. **Measured entities, in coordination with their measure implementation vendors, should conduct a feasibility assessment when first implementing the measure** to determine the internal and external capabilities needed to capture and extract the AI-derived data and share these findings with the

program owner and measure developer. Understanding what was required by each measured entity, particularly for those that are new to using AI and/or have limited resources, will be essential information to share. **Measure implementation vendors should apply their expertise and experience to help measured entities determine the most efficient way for a measure utilizing an AI-derived component to be implemented** within the entity's respective organization/institution. Throughout this process, program owners, measured entities, and measure implementation vendors should aim to reduce burden by identifying opportunities to align implementation practices across measured entities.

STRATEGY 4: ASSESS REGULARLY FOR UNINTENDED CONSEQUENCES, INCLUDING BIAS

TEP Recommendations

- Measure developer:
    - Stratifies component performance results by patient characteristics (included in quality measure AI model summary label)
    - Reviews stratified performance results from measured entities and measure implementation vendors to assess for bias and other unintended consequences, and when feasible, adjust the measure to mitigate these issues
- Program owner:
    - Reviews stratified performance results from measure developer and measured entities to assess for bias and other unintended consequences in outputs from the AI-derived component
- Measured entity with assistance from the measure implementation vendor:
    - Stratifies component performance results by patient characteristics and reports to program owner and measure developer

Stratification of AI-Derived Component Performance Results

Measure developers and measured entities, with assistance from measure implementation vendors, should stratify performance results for the AI-derived component by patient characteristics. Program owners and measure developers should analyze the differences in performance scores across patient subgroups to identify potential unintended consequences, including bias. Measures are currently assessed for their potential to encourage high-quality care for all with measure developers providing data on whether there are differences in performance scores across patient characteristics. Transparency around the differences in performance scores, as well as the transparency of the data used to develop and test the component (as discussed above in *Strategy 1*), are critical when assessing for potential bias.

Because AI-derived components may perpetuate inequities based on underlying systemic biases in care patterns and data collection informing the data used to develop the components, as described in the *Background* section above, measures with an AI-derived component should be assessed for equity and fairness throughout the development and implementation process. Because AI is an evolving technology, the TEP also underscored evaluating the component for other unintended consequences (e.g., potential impacts on patient safety) both at the patient and measured entity level.

While the TEP emphasized the importance of stratifying component performance results, members acknowledged that stratification may not currently be feasible or may be unreliable due to the lack of

sufficient data or small patient populations. For example, certain patient characteristics may be underreported or inconsistently captured in EHRs and/or administrative data, and data may not fully reflect the diversity of patient populations. This problem may be exacerbated by poor quality for existing data or data "missing not at random" (i.e., the likelihood of missing data is related to the unobserved values itself, suggesting that missing data differ systematically from observed values).[51]

Additionally, the TEP noted that the infrastructure to support comprehensive stratification, such as standardized data collection and interoperability, is still under development. The TEP recognized these limitations while highlighting how transparency and information about the performance of AI-derived components across different patient populations provide critical insights into potential unintended consequences. One TEP member suggested that additional validation analyses, such as sensitivity tests of subgroups, may further assist measure developers in understanding the potential biases in AI-derived components. For example, a predictive model developed using unrepresentative development data may misestimate the outcome of interest if implemented in a data set with different demographic characteristics than those used for development. As further described in *Strategy 6*, the TEP also suggested establishing a centralized reporting system to document and monitor unintended consequences or failures of the AI-derived component.

STRATEGY 5: PRIORITIZE ONGOING MONITORING AND MAINTENANCE

TEP Recommendations

- Measure developer (the measure steward will likely play a significant role in executing these recommendations):
    - Describes a monitoring and maintenance plan for the AI-derived component (included in quality measure AI model summary label)
    - Monitors results of the AI-derived component on an ongoing basis using feedback from the program owner, measured entities, and measure implementation vendors, including assessing for potential gaming, bias, and unintended consequences
    - Maintains the AI-derived component through regular updates and shares information with the program owner, measured entities, and measure implementation vendors about updates
- Program owner:
    - Develops a cadence and process for measured entities and measure implementation vendors to regularly assess component performance against performance standards
    - Collates and shares aggregate performance results and validation data, as available, stratified by patient characteristics, with measure developer
    - Monitors results of the AI-derived component on an ongoing basis, including assessing for potential gaming, bias, and unintended consequences
    - Shares periodic feedback from measured entities and measure implementation vendors, as available, with measure developer about AI-derived component
- Measured entity with assistance from the measure implementation vendor:
    - Provides feedback on the AI-derived component, performance results, and validation data (latter two items stratified by patient characteristics) to the program owner and measure developer

  - Assesses component performance against performance standards according to frequency established by program owner
  - Shares information on the timing and types of ongoing monitoring conducted for the measure and its component
  - Monitors and validates regularly the AI-derived component
- Measure implementation vendor:
  - Leverages their experience and expertise to assist measured entities to effectively monitor and maintain performance of the measure and its AI-derived component

Implementing a Monitoring and Maintenance Plan

Measure developers should have a monitoring and maintenance plan for the AI-derived component, including a description of the process and frequency for retesting the component. This plan will allow developers to understand how the component is performing across a broad range of measured entities outside of the data set in which they developed the component. The learnings can also help developers reduce data collection burden and facilitate alignment of AI-derived components assessing the same clinical concept. The importance of monitoring and maintenance is a central feature of AI governance and guidance frameworks due to growing evidence of model drift (i.e., the gradual degradation of model performance due to changes in data patterns and/or the relationship between input and output variables). Model drift can impact algorithms using a variety of methods.[52,53] Assessing for model drift is crucial because degradation can reduce the accuracy and consistency of a component's performance. It is also important to monitor model improvement because there is the potential for ongoing algorithmic development to positively impact performance of the underlying AI method. Additionally, developers should periodically review published evidence related to algorithm development and/or performance to verify that the model remains aligned with current evidence.

Measure developers should maintain the AI-derived component through regular updates. Several TEP members suggested the maintenance should take place annually and consider feedback from measured entities, measure implementation vendors, and program owners. The component should also be revalidated after any major clinical or data infrastructure change. **If the developer updates the component, they should share information with program owners, measured entities, and measure implementation vendors,** which is consistent with the current process for quality measures.

Monitoring Against and Updating Performance Standards

Program owners should develop a cadence and process for measured entities and measure implementation vendors to assess component performance against performance standards over time to ensure accuracy and consistency of the measure and its AI-derived component.[54] **Program owners should collate and share aggregate performance results and validation data from measured entities and measure implementation vendors, as available, with measure developers** so that developers can assess and interpret performance results in order to update the AI-derived component.

Data from the AI-derived component needs to be monitored and validated over time and it is important for program owners, measured entities, and vendors to share performance scores, validation data (including data stratified by patient characteristics), and the number and type of reporting entities with the measure developer. Measure developers should assess and interpret performance results and validation data to update (maintain) the component. Such coordination and communication between

entities allows the measure developer to continue to improve the component's performance and generalizability across a range of measured entities.

Program owners and measure developers should monitor results of the AI-derived component on an ongoing basis to ensure the component meets performance standards and assess for potential gaming and unintended consequences. Program owners and measure developers should review performance results on a regular schedule (e.g., annually) to determine if updates are needed to the performance standards or the AI-derived component, and to assess whether evidence of gaming or unintended consequences is emerging.

Measured entities and measure implementation vendors should also regularly monitor and validate the AI-derived component, using strategies such as random sampling and/or human review of component performance. The TEP noted that, similar to current processes used to monitor human chart abstractors, there needs to be oversight and evaluation at the local level to confirm the component is performing adequately. TEP members recognized that this level of monitoring may be more extensive than what is currently practiced for traditional quality measures; however, because of the potential for model degradation and drift, regular monitoring of the component is essential. If measured entities and vendors are unable to implement ongoing monitoring practices, the TEP advised that program owners may want to compare component results from measured entities and vendors that are not performing ongoing monitoring against a third-party data set.

Measured entities and measure implementation vendors should be transparent on the timing and types of ongoing monitoring they conduct for the measure and its component. The TEP acknowledged that while ongoing monitoring and maintenance is needed to establish trust in AI-enabled measure scores, there is considerable variability among measured entities' ability to access the resources and personnel needed to successfully operationalize monitoring and maintenance activities. TEP members emphasized the need to balance the risk of disadvantaging less-resourced measured entities with the potential harm that could arise if components are not adequately monitored.

STRATEGY 6: SUPPORT AN ECOSYSTEM THAT ENABLES INFORMATION SHARING AND FEEDBACK ACROSS KEY ACTORS

TEP Recommendations

- Measure developer:
    - Needs access to performance results, feedback on ongoing monitoring, feasibility, issues, and lessons learned from measured entities and measure implementation vendors
- Program owner:
    - Needs access to performance results, feedback on ongoing monitoring, feasibility, issues, and lessons learned from measured entities and measure implementation vendors
- Measured entity with assistance from the measure implementation vendor:
    - Needs mechanism for sharing performance results, feedback on ongoing monitoring, feasibility, issues, and lessons learned with measure developer and program owner
    - Needs mechanism for submitting questions and providing feedback on emerging limitations and risks to measure developer
    - Needs information on updates to AI-derived component

Establishing a Feedback Loop Process

A feedback loop process that supports sharing of information across key actors throughout the measure lifecycle needs to be established. As described in Strategies 1-5, measure developers and program owners need to receive periodic updates from measured entities and measure implementation vendors, including information on performance results, feedback on ongoing monitoring, feasibility, issues, and lessons learned from measured entities and measure implementation vendors. To reduce measure cacophony and duplicative data collection efforts, this process could also allow key actors to provide input on ways to align specifications for different AI-enabled measures assessing the same clinical concept. However, this process does not currently exist in quality measurement and requires significant collaboration between key actors to establish and maintain.

One suggestion provided by the TEP was developing a centralized website or communication channel where program owners, measured entities, and measure implementation vendors can easily reference quality measure AI model summary label information for the AI-derived component and submit any related questions or issues to the measure developer. The TEP recognized that the measure developer's ability to establish and maintain this site will vary based on the type of developer and their readily available resources. In addition, the TEP recommends establishing a centralized reporting system—similar to pharmacovigilance systems—for measure developers, program owners, measured entities, and vendors to document, report, and monitor unintended consequences of the AI-derived component. This system would support transparency and build a shared understanding of model limitations and emerging risks.

Roadmap for Implementing TEP Recommendations

To assist measure developers, program owners, measured entities, and measure implementation vendors with implementing these recommendations, NQF and the TEP translated the recommendations into specific actions and responsibilities across five key steps in the measure lifecycle: (1) development and testing, (2) selection, (3) preparation for implementation, (4) implementation across entities, and (5) monitoring and maintenance (Table 5).

Table 5. Description of the Five Steps of the Measure Lifecycle

Key Step	Description
1. Development and Testing	This step involves identifying gaps in measurement and the need for a measure; conceptualizing and sufficiently detailing the measure calculation/specification; and assessing the measure for feasibility, usability, and scientific acceptability. The measure developer considers the advantages and disadvantages of an AI-derived component. If measure developers choose to include an AI-derived component, they should develop, test, and describe it in the quality measure AI model summary label.

Key Step	Description
2. Selection	This step begins with a fully developed and tested measure, including its AI-derived component, as a program owner considers using the measure in an accountability program. The program owner reviews measure information (i.e., determines if the measure is important, feasible, reliable, and valid) and the development and testing details, (including the performance of the AI-derived component) to assess the measure's readiness and appropriateness for the program.
3. Preparation for Implementation	This step begins after a program owner selects a measure for use in an accountability program. Activities ensure the measure produces reliable and valid results, and its AI-derived component is generalizable and performs accurately across measured entities before scores are used for accountability decisions (e.g., public reporting, financial incentives). This step will vary by program but may include piloting the measure and its AI-derived component, testing the measure and its AI-derived component, implementing pay for reporting, a measure dry run, and/or voluntary reporting.
4. Implementation Across Entities	This step begins after the program owner determines scores from an AI-enabled measure are accurate, appropriate, and ready to use for accountability decisions. The implementation process involves scaling the measure and its AI-derived component across a large number of measured entities and using results for accountability decisions.
5. Monitoring and Maintenance	This step involves monitoring and updating the measure and its AI-derived component as needed. If the measure and/or its AI-derived component change significantly, the measure may need to re-enter the process at one of the preceding steps.

As illustrated in Table 5, the TEP defined a critical step, Step 3, Preparation for Implementation, between a program owner selecting a measure with an AI-derived component for a program and the program owner using the measure results (e.g., adjusting payment or publicly reporting the scores). This step already exists for many accountability programs and allows the program owner to assess whether a measure will perform as expected across measured entities or produce unexpected consequences before using the results, but is not consistently implemented across all programs.

The amount and type of testing performed during the preparation step is informed by prior testing conducted by the measure developer, because it can be highly variable. For example, a program owner may choose to forego the preparation step if a measure developer has conducted national testing of the measure and its AI-derived component or used a nationally representative data set for development and testing of the component. Therefore, how the program owner best prepares a measure for implementation in their program will vary. It may involve, for example, piloting the AI-derived component with a limited number of measured entities or all potential measured entities, asking measured entities to voluntarily report the measure, implementing pay for reporting, and/or collecting data from all measured entities but not using the results for decision-making.

The TEP conceptualized these five steps as a measure lifecycle. Figure 3 highlights the key activities that should occur during the steps.

Figure 3. Five-Step Roadmap for AI-Enabled Quality Measures

The figure below depicts the five-step roadmap for developing and testing, selecting, implementing, and monitoring and maintaining AI-enabled quality measures. Each step highlights key activities grouped by lead actor (i.e., program owner, measure developer, measured entity, measure implementation vendor). However, successful execution of these activities will require collaboration across all key actors. The roadmap begins with (1) development and testing, where measure developers assess feasibility of implementation, rigorously test the measure and its component, design the measure to mitigate for potential gaming, and complete the quality measure AI model summary label. During (2) selection and (3) preparation for implementation, program owners use the quality measure AI model summary label to determine the measure's appropriateness for use in a program. During these steps, program owners also conduct a feasibility assessment, including testing with measured entities, and establish performance standards prior to widespread implementation. During (4) implementation across entities and (5) monitoring and maintenance, measured entities and measure implementation vendors implement, test, and tune the component, as needed; and validate performance against established standards. Entities report these results, and measure developers conduct further maintenance of the component based on feedback.

MEASURE DEVELOPER:

- Evaluates feasibility of the component's implementation
- Rigorously tests the measure and component
- Completes the quality measure AI model summary label
- Designs measure to mitigate potential gaming

PROGRAM OWNER:

- Leverages the quality measure AI model summary label information to determine appropriateness during the selection process
- Performs a feasibility assessment
- Establishes performance standards for the component

MEASURE DEVELOPER:

- Maintains component based on results and feedback from measured entities

PROGRAM OWNER:

- Monitors performance results from measured entities

MEASURED ENTITY AND MEASURE IMPLEMENTATION VENDOR:

- Implement, test, and tune the component as needed
- Validate component performance against performance standards and report results
- Monitor component

ROADMAP TABLES FOR THE FIVE KEY STEPS

For each of the steps, the TEP developed roadmap tables (Tables 6-10) that define the actions for: (1) measure developers, (2) program owners, (3) measured entities, and (4) measure implementation vendors. Some tasks are relevant across more than one of these actors. Therefore, to clarify and differentiate roles across key actors, the TEP applied the RACI framework[1](see box) to define each actor's relative role (responsible, accountable, consulted, or informed) in executing recommendations. It is important to note that, although the TEP's discussions primarily focused on the AI-derived component of a measure, many of the tasks described in the recommendation tables below may apply to the entire measure.

Responsible entities: Tasked with successful completion and/or implementation of assigned recommendations in the process. These entities follow guidance established by accountable entities to ensure effective development, selection, and implementation of measures using AI methods.

Accountable entities: Tasked with ensuring that assigned recommendations in the process are implemented as intended. These entities provide background and expectations for successful development, selection, and implementation of measures using AI methods.

Consulted entities: Tasked with providing feedback and input to support recommendations in the process. These entities may not be directly responsible/accountable, but they have a stake in the outcomes and can help inform steps in the process.

Informed entities: These entities do not assume specific tasks, because they are not decision makers or directly responsible/accountable. However, they may be indirectly involved and therefore should be informed about developments in the process.

TABLE 6. ROADMAP STEP 1: DEVELOPMENT AND TESTING

Key Actor	TEP Recommendations
Measure Developer	**Accountable for:** • Considering the advantages/disadvantages of including an AI-derived component in the measure and choosing whether to use an AI-derived component • Designing the measure and its AI-derived component to avoid the potential for gaming • Testing the measure according to existing consensus-based criteria, while also conducting specific testing of the AI-derived component • Conducting evaluation of the AI-derived component performance that meets current industry standards • Developing the component using separate measured entities or data sets from those used for testing • Stratifying testing results by patient characteristics • Identifying and describing the resources needed by measured entities to implement the AI-derived component • Describing a monitoring and maintenance plan for the AI-derived component • Completing a quality measure AI model summary label according to the standardized template
Program Owner	**Consulted about:** • Information in the quality measure AI model summary label **Informed about:** • Measure developer's approach to minimizing the potential for gaming of the measure and its AI-derived component
Measured Entity	**Informed about:** • Quality measure AI model summary label content
Measure Implementation Vendor	**Informed about:** • Quality measure AI model summary label content

TABLE 7. ROADMAP STEP 2: SELECTION

Key Actor	TEP Recommendations
Measure Developer	**Responsible for:** • Providing a completed quality measure AI model summary label • Providing a configuration file for the AI-derived component and other information that aids implementation of the component

Key Actor	TEP Recommendations
Program Owner	**Accountable for:** • Establishing selection criteria that include consideration of quality measure AI model summary label information • Confirming measure developer provides the completed quality measure AI model summary label • Consulting the quality measure AI model summary label to evaluate the component's performance and its appropriateness and feasibility for the program, program setting, and measured entities
Measured Entity	**Consulted about:** • Technical approach for the component, including its potential impact on workflow, the feasibility of its data inputs, resources needed to implement the component, and face validity **Informed about:** • Quality measure AI model summary label content
Measure Implementation Vendor	**Consulted about:** • Technical approach for the component, including its potential impact on workflow, the feasibility of its data inputs, resources needed to implement the component, and face validity **Informed about:** • Quality measure AI model summary label content

TABLE 8. ROADMAP STEP 3: PREPARATION FOR IMPLEMENTATION

Key Actor	TEP Recommendations
Measure Developer	**Responsible for:** • Providing guidance to measured entities and measure implementation vendors about testing they should conduct locally to validate output from the AI-derived component • Providing guidance to program owners about appropriate performance standards • Assessing and interpreting stratified performance results and validation data from measured entities and measure implementation vendors to update the component, and assess for potential gaming, bias, and unintended consequences • Addressing issues or questions from measured entities and measure implementation vendors **Informed about:** • Feedback from measured entities and measure implementation vendors on implementation of the component

Key Actor	TEP Recommendations
Program Owner	**Accountable for:** • Determining if additional testing of the AI-derived component is needed based on testing completed by measure developer • Performing a feasibility assessment of the AI-derived component, as needed, prior to widespread implementation • Establishing performance standards against which measured entities compare component results • Ensuring participating measured entities and measure implementation vendors compare performance of the component against performance standards • Monitoring performance results of the component on an ongoing basis, including reviewing stratified results to assess for potential gaming, bias, and unintended consequences
Measured Entity	**Responsible for:** • Assessing and reporting on the feasibility of implementing the component • Implementing the component with patient data and testing the component's performance • Comparing results of the component against the performance standards set by the program owner • Tuning the component, as needed, to meet performance standards and retesting performance after tuning • Reporting performance results stratified by patient characteristics and steps taken to tune the component to program owner and measure developer
Measure Implementation Vendor	**Responsible for:** • Assessing and reporting on feasibility of implementing the component • Supporting measured entities to validate implementation • Comparing performance of the component against established performance standards

TABLE 9. ROADMAP STEP 4: IMPLEMENTATION ACROSS ENTITIES

Key Actor	TEP Recommendations
Measure Developer	**Responsible for:** • Assessing and interpreting stratified performance results and validation data from measured entities and measure implementation vendors to update the component, and assess for potential gaming, bias, and unintended consequences **Consulted about:** • Issues or questions from measured entities and measure implementation vendors **Informed about:** • Feedback on implementation of the component
Program Owner	**Accountable for:** • Updating performance standards based on results from Step 3 • Establishing performance standards against which measured entities compare component results • Ensuring participating measured entities and measure implementation vendors compare performance of the component against performance standards • Monitoring performance results of the component on an ongoing basis, including reviewing stratified results to assess for potential gaming, bias, and unintended consequences • Sharing feedback on the component from measured entities and measure implementation vendors with measure developers
Measured Entity	**Responsible for:** • Conducting a feasibility assessment to determine internal and external capabilities needed to capture and extract the component • Implementing the component with patient data and testing the component's performance • Comparing results of the component against the performance standards set by the program owner • Tuning the component, as needed, to meet performance standards and retesting performance after tuning • Reporting performance results stratified by patient characteristics and steps taken to tune the component to program owner and measure developer • Providing feedback to the program owner and measure developer, including resources and internal and external capabilities required
Measure Implementation Vendor	**Responsible for:** • Assessing and reporting on feasibility of implementing the component • Supporting measured entities to validate implementation • Comparing performance of the component against established performance standards

TABLE 10. ROADMAP STEP 5: MONITORING AND MAINTENANCE

Key Actor	TEP Recommendations

Measure Developer	**Responsible for:**
	• Implementing monitoring and maintenance plan for the AI-derived component, including:
	  ○ Assessing and updating the component regularly
	  ○ Retesting the AI-derived component for ongoing performance
	  ○ Assessing and interpreting performance results and validation data from measured entities and measure implementation vendors to update the component
	  ○ Assessing for potential gaming, bias, and unintended consequences and, when feasible, adjusting the measure to mitigate these issues
	  ○ Informing program owners, measured entities, and measure implementation vendors about updates to the component
	Consulted about:
	• Issues or questions from measured entities and measure implementation vendors
	Informed about:
	• Feedback on implementation of the component

Program Owner	**Accountable for:** • Sharing feedback from measured entities and measure implementation vendors with measure developers about the AI-derived component • Developing cadence and process for measured entities to regularly assess component performance against performance standards • Collating and sharing aggregate performance results and validation data, stratified by patient characteristics, with measure developers • Monitoring results of the component on an ongoing basis, including assessing for potential gaming, bias, and unintended consequences **Informed about:** • Updates made to the component • Feedback on the component, performance results, and validation data (the latter two stratified by patient characteristics)
Measured Entity	**Responsible for:** • Providing feedback on the component, performance results, and validation data (latter two stratified by patient characteristics) • Assessing component performance against performance standards according to the frequency established by the program owner • Providing information on the timing and types of ongoing monitoring • Monitoring and validating regularly the AI-derived component **Informed about:** • Updates made to the component

Key Actor	TEP Recommendations

Measure Implementation Vendor	**Responsible for:**
	• Assisting with monitoring performance of the AI-derived component, including:
	  ○ Working with the measured entity to advise on the ongoing feasibility of implementation
	  ○ Providing feedback on the component, performance results, and validation data (latter two stratified by patient characteristics)
	  ○ Assessing component performance against performance standards according to frequency established by program owner
	  ○ Providing information on the timing and types of ongoing monitoring
	  ○ Monitoring and validating regularly the AI-derived component
	  ○ Sharing lessons learned with the program owner, measured entities, and measure developer
	Informed about:
	• Updates made to the component

Emerging Topics

In addition to the recommendations contained in the five-step process, the TEP identified several emerging topics for consideration when developing, selecting, and implementing AI-enabled quality measures. The TEP acknowledged the importance of finding resolutions for these issues; however, because the use of and considerations around AI are rapidly evolving, the TEP did not develop formal recommendations on these topics. The TEP advises that these issues need further consideration, and future recommendations related to these topics should encourage the trustworthy use of AI in quality measurement while advancing the types of innovative measurement that AI methods allow. The identified emerging topics are:

- Sharing code and weights and/or proprietary details for the AI-derived component
- Validating the AI-derived component with a third-party evaluator and/or reference data set
- Allowing measured entities to select their own AI method

SHARING CODE AND WEIGHTS AND/OR PROPRIETARY DETAILS FOR THE AI-DERIVED COMPONENT

Several TEP members agreed that while it would be ideal for measure developers to share details about the AI-derived component of a measure (e.g., programming code, weights) to promote transparency, it may not always be possible. The amount and type of information a measure developer can share will depend on the type of AI used (e.g., NLP, LLM) and whether that algorithm is proprietary. In situations where measure developers cannot provide programming code or weights or descriptive statistics about the data sets used to develop the component, information about the testing data and results may be even more critical.

Some TEP members proposed that measure developers provide technical details about the component to the greatest extent possible. However, other members noted that understanding the data used to train the component rather than the algorithm itself would provide valuable information and therefore the exact weights and code would not be required. The TEP advised that it would be preferable to encourage measure developers to allow groups to query that component because it would provide better insights into whether it captures the clinical concept as intended and performs as expected.

The real-world implementation results—including performance of the component against defined performance standards across a diverse group of measured entities—will be essential, especially in instances where the component details are not transparent. Several TEP members proposed that there could be a centralized library like the Value Set Authority Center, which is maintained by the National Library of Medicine, where measure developers could share and update their AI-derived components but that sharing components should not be a requirement.

VALIDATING THE AI-DERIVED COMPONENT WITH A THIRD-PARTY EVALUATOR AND/OR REFERENCE DATA SET

The TEP discussed the potential value of external validation (i.e., validation by an entity other than the measure developer) of the results of the AI-derived component against a gold standard, with several members suggesting the possibility of centralizing the validation process through a neutral third-party evaluator because this could broaden access to generalizable data sets and support an independent assessment of component results. The TEP did not develop any recommendations on this external validation concept because there is not yet a gold standard against which to assess AI components and these evaluators do not yet exist. Additionally, some TEP members raised concerns about the establishment of a "cottage industry" to validate AI-derived components and the cost of having third-party evaluators review all AI-derived components for quality measures.

Several TEP members also suggested the need for a third-party reference data set in which measured entities could evaluate the AI-derived component against performance standards to verify that measured entities are producing expected results prior to applying the measure to their patient population. Reference data sets would also help entities identify and address any sources of error in the component prior to implementation. Validation against a third-party data set could provide a higher degree of confidence in and credibility to the measure and resulting data.

One TEP member noted that Leapfrog's Computerized Provider Order Entry (CPOE) Evaluation Tool could serve as a model for the non-public data set that measured entities could use to validate how the AI-derived component is functioning in their data. As part of Leapfrog's program, hospitals implement the data set maintained by Leapfrog, which includes medication orders and test patients created based on published literature of known medication errors. Hospitals run the data set through the CPOE system and evaluate performance based on the degree to which the system produces the desired results.[55] The TEP advised that reference data sets could be informed by a sensitivity analysis of the component which assesses how variations in inputs affect the outputs, as well as an understanding of how data varies across measured entities.

These validations could continue throughout the implementation process, as needed. The TEP envisioned a collaborative approach between the program owner, measured entities, measure

implementation vendors, and measure developer where results from the component are validated and feedback is provided to further improve performance of the component and data quality.

One TEP member cautioned that implementing reference data sets may create burden for entities because it takes time and resources to run, process, and certify the data set. Some TEP members questioned the feasibility of developing large data sets against which these components could be validated, noting that the data quickly become outdated. In addition to the mixed feedback about the use of reference data sets, the TEP had mixed reactions about constructing a reference data set with synthetic or real patient data. Several TEP members suggested it would be more feasible to use synthetic data because aggregating real patient data, even when deidentified, raises privacy and security concerns. Centralizing large volumes of patient data in a single location can increase vulnerability to cyber threats, underscoring the need for caution when considering data sharing. Several TEP members supported the use of synthetic data, noting that the quality of synthetic data generation is expected to improve in the future. However, other TEP members argued that synthetic data may not capture the complexity of real clinical notes and has historically not been as effective for training or evaluation of AI models. A few TEP members proposed the use of reference data sets that are constructed using a mixture of synthetic and real data.

Locked: "A model that provides the same output each time the same input is applied to it and does not change with use, as its parameters or configuration cannot be updated."[56]

Continual Machine Learning: "The ability of a model to adapt its performance by incorporating new data or experiences over time while retaining prior knowledge/information… In contrast to a locked model, a continual machine learning model has a defined learning process to change its behavior."[56]

The TEP also considered how the validation process may differ for components based on "locked" or "continual machine learning" models (see box) used in AI-enabled quality measures.[56]The TEP noted that components may evolve across measured entities and over time for several reasons; for example, entities may tune the component to account for local contexts, the developer of the model underlying the AI-derived component may update the model (e.g., with LLMs), or the component may learn in each local context if it is continually learning.

As AI models evolve and underlying algorithms are updated, it remains unclear as to when updates to a component constitute a significant enough change to warrant updated validation. Several TEP members did not want to limit quality measures to only using locked components because they thought this would limit measured entities from tuning components and prevent the use of LLMs, while others cautioned that it may be difficult to compare scores across measured entities when quality measures use tuned or learning components.

For both locked and continual machine learning components, the TEP advised monitoring over time, but did not reach consensus on when updated validation is necessary, underscoring the need for future guidance to confirm that updates to AI-derived components do not compromise the integrity of measure scores. Third-party evaluation and/or reference data sets may play a role in monitoring and ongoing validation of components, whether they are locked or continually learning.

ALLOWING MEASURED ENTITIES TO SELECT THEIR OWN AI METHOD

Some TEP members suggested that the five-step process outlined in this guidance should allow for measured entities to choose their AI method, regardless of the AI-derived component originally developed and tested by the measure developer, as long as the entities can demonstrate their AI model meets the measure's intent and meets established performance standards. For example, a measured entity may determine that they can achieve greater accuracy using an LLM to collect data compared to the NLP provided by the measure developer. Several members suggested that application of a universal AI model could not occur across all measured entities without the entities needing to tune the component. Therefore, it is important to allow entities flexibility in selecting an AI method that optimizes the measure concept and works best for their respective organization/institution.

The TEP further acknowledged the challenge in expecting measured entities to implement AI-derived components exactly as specified by measure developers. Several TEP members agreed that while allowing entities to choose their own AI method offers flexibility, it also raises concerns about maintaining consistency and accuracy of implementation across entities, which will compromise benchmarks and cross-site comparisons within accountability programs. For example, some measured entities may not have the resources needed to implement LLMs, highlighting a potential disparity between entities using NLP versus LLMs for quality measurement. These members emphasized that allowing the use of different AI methods could result in inconsistencies, and suggested that if entities are allowed this flexibility, they should be subject to more frequent assessments of the AI-derived component against performance standards.

The TEP also noted that these differences could impact the ability to compare performance scores across measured entities, highlighting the need for future work to identify strategies that could mitigate this challenge because diverse AI methods are increasingly adopted across different measured entities. One TEP member suggested an alternate approach whereby program owners use the quality measure AI model summary label to establish criteria that would help inform measured entities and measure implementation vendors as to which AI models are appropriate to apply to the measure. Additionally, some TEP members advised that evaluation against a reference data set is particularly important for verifying consistency of results across measured entities that are implementing the component using different AI methods and there could be value in having the component tested against non-public data sets that are developed for validating the component and not used to train the model.

Conclusion

The use of AI in quality measures shows promise in reducing measurement burden while allowing significant development in areas that previously have been difficult or burdensome to measure. As the development and use of AI-enabled quality measures continues to advance, it is essential to establish and maintain guidance on how to effectively develop, select, and implement quality measures that use AI methods for use in accountability programs.

The recommendations outlined in this framework are an initial step in establishing this guidance for program owners, measure developers, measured entities, and measure implementation vendors. The TEP recognized that the use of AI in healthcare is rapidly evolving towards more complex AI methods

and to keep pace with these changes, it is critical that the guidance and recommendations outlined in this report are assessed and updated over time.

The TEP flagged several important considerations related to the resources necessary to implement AI-enabled quality measures. For instance, program owners and measured entities will vary in their ability to implement AI-derived components based on their financial, staffing, and computing resources. As an example, the TEP noted that a larger academic healthcare system may be more able to access the resources needed to contract with measure implementation vendors to effectively implement measures and their AI-derived components, compared to a smaller community-based hospital. Because of this variability in the healthcare landscape and the evolving nature of AI, the TEP emphasized that their recommendations reflect an ideal ecosystem in which key actors have the resources needed to feasibly develop, select, and implement AI-enabled quality measures. However, not all key actors will have the ability to effectively operationalize all recommended strategies.

Finally, the TEP recognized that while the recommendations in this report apply to developing, selecting, and implementing measures that use AI methods in regional and national accountability programs, they may have implications for the development and use of measures for other purposes, including quality improvement activities. As the use of AI-enabled quality measures advances, it will be important to consider adapting and applying these recommendations to quality measures used for purposes outside of accountability programs.

The recommendations may also have implications for other parts of the quality measure development and implementation process (e.g., consensus-based entity review for endorsement, pre-rulemaking review). Overall, the recommendations included in this framework are intended to be a first step in establishing governance over the use of AI in quality measurement. Adaptation will be necessary over time to account for factors such as emerging AI methods, and cost implications, as well as additional use cases of quality measures that use AI methods.

References

1. Miranda D, et al. *What is a RACI chart?*; 2024. https://www.forbes.com/advisor/business/raci-chart/. Accessed July 1, 2025
2. Mitchell M, et al. *Model cards for model reporting*. ACM Digital Library; 2019. https://dl.acm.org/doi/10.1145/3287560.3287596. Accessed July 1, 2025
3. European Information Technologies Certification Academy (EITCA). *What are weights and biases in AI?*; 2023. https://eitca.org/artificial-intelligence/eitc-ai-gcml-google-cloud-machine-learning/introduction/what-is-machine-learning/explain-weights-and-biases/. Accessed 8/27/25
4. Garvin JH, et al. *Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure*. Journal of the American Medical Informatics Association; 2012; 5. 10.1136/amiajnl-2011-000535. Accessed July 1, 2025
5. Mehrotra A, et al. *Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures*. Gastrointestinal Endoscopy; 2012; 6. 10.1016/j.gie.2012.01.045. Accessed July 1, 2025
6. Greenberg JO, et al. *Journal of the American Medical Informatics Association : JAMIA*. 2013;20:e97-e101

7. Lee RY, et al. *Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome*. JAMA Network. 10.1001/jamanetworkopen.2023.1204. Accessed July 1, 2025
8. Bannett Y, et al. *Applying large language models to assess quality of care: monitoring ADHD medication side effects*. American Academy of Pediatrics; 2025. 10.1542/peds.2024-067223. Accessed July 1, 2025
9. Sippo DA, et al. *Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing*. Journal of Digital Imaging; 2013. 10.1007/s10278-013-9616-5. Accessed July 1, 2025
10. Centers for Medicare & Medicaid Services (CMS). *Pre-rulemaking MUC lists and recommendation reports*; 2024. https://mmshub.cms.gov/measure-lifecycle/measure-implementation/pre-rulemaking/lists-and-reports. Accessed July 1, 2025
11. Adams L, et al. *Artificial intelligence in health, health care, and biomedical science: an AI code of conduct principles and commitments discussion draft*. National Academy of Medicine (NAM); 2024. https://nam.edu/wp-content/uploads/2024/04/Artificial-Intelligence-in-Health-Health-Care-and-Biomedical-Science_final_4.8.24V2.pdf. Accessed July 1, 2025
12. Office of the National Coordinator for Health Information Technology (ONC). *Health data, technology, and interoperability: Certification program updates, algorithm transparency, and information sharing*; 2024. https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and. Accessed July 1, 2025
13. U.S. Food and Drug Administration (FDA). *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)*; 2019. https://www.fda.gov/media/122535/download?attachment. Accessed July 1, 2025
14. U.S. Food and Drug Administration (FDA). *Artificial intelligence (AI) and machine learning (ML) in medical devices*; 2020. https://www.fda.gov/media/151482/download. Accessed July 1, 2025
15. U.S. Food and Drug Administration (FDA). *Artificial intelligence and machine learning in software as a medical device*; 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed July 1, 2025
16. Partnership for Quality Measurement (PQM). *Endorsement and maintenance (E&M) guidebook*; 2025. https://p4qm.org/sites/default/files/2025-08/Del-3-6-Endorsement-and-Maintenance-Guidebook-OP2-508.pdf. Accessed August 25, 2025
17. National Quality Forum (NQF). *Measure evaluation criteria and guidance for evaluating measures for endorsement*; 2023. https://cms.qualityforum.org/Measuring_Performance/Submitting_Standards/2021_Measure_Evaluation_Criteria_and_Guidance.aspx. Accessed July 1, 2025
18. Centers for Medicare & Medicaid Services (CMS). *Blueprint measure lifecycle*; 2024. https://mmshub.cms.gov/blueprint-measure-lifecycle-overview. Accessed July 1, 2025
19. Partnership for Quality Measurement (PQM). Available at: https://p4qm.org/measures/3749e. Accessed July 1, 2025
20. American Medical Association (AMA). *Augmented intelligence development, deployment, and use in health care*; 2024. https://www.ama-assn.org/system/files/ama-ai-principles.pdf. Accessed July 1, 2025
21. IBM. *What is predictive AI?*; 2024. https://www.ibm.com/think/topics/predictive-ai. Accessed July 1, 2025
22. IBM. *Artificial intelligence*; 2022. https://www.ibm.com/design/ai/basics/ai/. Accessed July 1, 2025
23. U.S. Government Accountability Office (GAO). *Health care quality measurement: the national quality forum has begun a 4-year contract with HHS*; 2010. https://www.gao.gov/assets/gao-10-737.pdf.

Accessed July 1, 2025
24. Centers for Medicare & Medicaid Services (CMS). *Measures management system (MMS) glossary*; 2025. https://mmshub.cms.gov/glossary. Accessed 8/21/25
25. Parker VJ, et al. *AI governance in health systems: aligning innovation, accountability, and trust*; 2024. https://healthpolicy.duke.edu/sites/default/files/2024-10/AI%20Governance%20in%20Health%20Systems.pdf. Accessed July 1, 2025
26. Obermeyer Z, et al. *Dissecting racial bias in an algorithm used to manage the health of populations*. Science; 2019. 10.1126/science.aax2342. Accessed July 1, 2025
27. Teeple S, et al. *Exploring the impact of missingness on racial disparities in predictive performance of a machine learning model for emergency department triage*. JAMIA Open; 2023. 10.1093/jamiaopen/ooad107. Accessed July 1, 2025
28. Rahman S, et al. *Generalization in healthcare AI: evaluation of a clinical large language model*; 2024. http://arxiv.org/pdf/2402.10965v2. Accessed July 1, 2025
29. Bohr A, et al. *The rise of artificial intelligence in healthcare applications*. Academic Press; 2020. 10.1016/B978-0-12-818438-7.00002-2. Accessed July 1, 2025
30. Yu K-H, et al. *Artificial intelligence in healthcare*. Nature Biomedical Engineering; 2018. 10.1038/s41551-018-0305-z. Accessed July 1, 2025
31. National Institute of Standards and Technology (NIST). *Artificial intelligence risk management framework (AI RMF 1.0)*; 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. Accessed July 1, 2025
32. Coalition for Health AI (CHAI). *Blueprint for trustworthy AI implementation guidance and assurance for healthcare*. https://assets.ctfassets.net/7s4afyr9pmov/4AXIWGIlcrjWDaW2ueTaRS/f98e5cb2528187635895cce6ba5ec309/Blueprint_for_Trustworthy_AI.pdf. Accessed July 1, 2025
33. The Light Collective. *AI rights for patients*; 2024. https://lightcollective.org/wp-content/uploads/2024/03/Collective-Digital-Rights-For-Patients_v1.0.pdf. Accessed July 1, 2025
34. U.S. Food and Drug Administration & Health Canada and the United Kingdom's Medicines and Healthcare Products Regulatory (MHRA). *Agency good machine learning practice for medical device development: guiding principles*; 2021. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. Accessed July 1, 2025
35. U.S. Food and Drug Administration (FDA). *Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations*; 2025. https://www.fda.gov/media/184856/download. Accessed July 1, 2025
36. National Academy of Medicine (NAM). Available at: https://doi.org/10.17226/29087. Accessed July 1, 2025
37. Consumer Technology Association (CTA). Available at: https://www.cta.tech/standards/ansicta-2116/. Accessed July 1, 2025
38. Consumer Technology Association (CTA). Available at: https://www.cta.tech/standards/ansicta-2090/. Accessed July 1, 2025
39. Siala H, et al. *Social science & medicine (1982)*. 2022;296:114782
40. Coalition for Health AI (CHAI). *The CHAI applied model card*; 2024. https://chai.org/draft-chai-applied-model-card/. Accessed July 1, 2025
41. Sendak MP, et al. *NPJ digital medicine*. 2020;3:41
42. Hernandez-Boussard Tina, Bozkurt Selen, Ioannidis John P A, Shah Nigam H. *MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care*; 2020. 10.1093/jamia/ocaa088. Accessed July 1, 2025
43. OpenAI Developer Community. *Understanding AI manipulation: a case study on the 'agitation'*

method; 2024. https://community.openai.com/t/understanding-ai-manipulation-a-case-study-on-the-agitation-method/594003. Accessed July 1, 2025
44. Draelos R. *Measuring performance: AUC (AUROC)*; 2022. https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/. Accessed July 1, 2025
45. Medium. *Unbalanced data? Stop using ROC-AUC and use AUPRC instead*; 2022. https://medium.com/data-science/imbalanced-data-stop-using-roc-auc-and-use-auprc-instead-46af4910a494. Accessed July 1, 2025
46. Medium. *Unraveling diagnostic performance metrics: a comprehensive guide to deriving sensitivity, specificity, PPV, and NPV from accuracy and confusion metrics*; 2023. https://medium.com/@codethulo/unraveling-diagnostic-performance-metrics-a-comprehensive-guide-to-deriving-sensitivity-9acfc9629b6e. Accessed July 1, 2025
47. Medium. *Confusion matrix, accuracy, precision, recall, F1 score*; 2019. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd. Accessed July 1, 2025
48. Medium. *MAE, MSE, RMSE, coefficient of determination, adjusted R squared — which metric is better?*; 2020. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. Accessed July 1, 2025
49. Weisburd D, et al. *Measuring association for scaled data: pearson's correlation coefficient*; 2021. https://link.springer.com/chapter/10.1007/978-3-030-47967-1_14. Accessed July 1, 2025
50. Gulem K. *What is a calibration curve?* https://dataconomy.com/2025/04/18/what-is-a-calibration-curve/. Accessed July 1, 2025
51. Snart H. *Understanding your data: Missing at random versus missing not at random*; 2021. https://blogs.sas.com/content/hiddeninsights/2021/08/16/understanding-your-data-missing-at-random-versus-missing-not-at-random/. Accessed August 21, 2025
52. Vela D, et al. *Temporal quality degradation in AI models*. Scientific Reports; 2022. 10.1038/s41598-022-15245-z. Accessed July 1, 2025
53. IBM. *What is model drift?*; 2024. https://www.ibm.com/think/topics/model-drift. Accessed July 1, 2025
54. National Association of ACOs (NAACOS). *Quality measure validation audit resource*; 2017. https://www.naacos.com/wp-content/uploads/2024/01/QMVAuditResource_v2.pdf. Accessed July 1, 2025
55. Leapfrog Group. *Prepare for CPOE tool*; 2024. https://www.leapfroggroup.org/survey-materials/prepare-cpoe-tool. Accessed July 1, 2025
56. Food and Drug Administration (FDA). Available at: https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource. Accessed July 1, 2025
57. Coalition for Health AI (CHAI). *Responsible AI guide*; 2024. https://assets.ctfassets.net/7s4afyr9pmov/6e7PrdrsNTQ5FjZ4uyRjTW/c4070131c523d4e1db26105aa51f087d/CHAI_Responsible-AI-Guide.pdf. Accessed July 1, 2025

Appendix A: Artificial Intelligence in Quality Measures Technical Expert Panel Members, Liaisons, Advisory Group Members, and NQF Staff

TECHNICAL EXPERT PANEL MEMBERS

Monica Agrawal, PhD, MS
Assistant Professor, Duke University
Co-Founder, Layer Health

Chandra Beasley, MBA, MHRM, MNSA, MPA, CLSSGB
Director of Information Technology, South Carolina Primary Health Care Association

Robert El-Kareh, MD, MPH, MS
Associate Chief Medical Officer, Transformation & Learning at UC San Diego Health

Mark Alan Fontana, PhD
Senior Director of Data Science, Hospital for Special Surgery
Assistant Professor, Weill Cornell Medical College

Rebecca Jacobson, MD, MS, FACMI
CEO and Founder, Astrata, Inc.

Laura D. Jantos, MBA, MHA, LFHIMSS
Patient Advocate and Consultant, LDJ Consulting, LLC

Rosemary Kennedy, PhD, RN, MBA, FAAN
Chief Health Informatics Officer, Connect America

Zhenqiu Lin, PhD
Senior Director of Healthcare Analytics, Yale New Haven Hospital Center for Outcomes Research and Evaluation

Charlotta Lindvall, MD, PhD, FAAHPM
Director of Clinical Informatics, Dana-Farber Cancer Institute, Boston

Vincent Liu, MD, MS
Chief Data Officer, The Permanent Medical Group
Senior Research Scientist, Kaiser Permanente Division of Research

Danielle A. Lloyd, MPH
Senior Vice President, Private Market Innovations and Quality Initiative, AHIP

Yuan Luo, PhD
Chief AI Officer, Northwestern University Clinical and Translational Sciences Institute

John Martin, PhD, MPH
Vice President, Data Science, Premier, Inc.

Eric Poon, MD, MPH, FACMI
Chief Health Information Officer, Duke University Health System

Paul Tang, MD, MS
Adjunct Professor, Clinical Excellence Research Center, Stanford University

Meghan Reading Turchioe, PhD, MPH, RN
Assistant Professor, Columbia University School of Nursing

Ben Wandtke, MD, MS
Vice Chair, Quality and Safety, Department of Imaging Sciences at University of Rochester Medical Center

LIAISONS

Laura Adams
Senior Advisor, National Academy of Medicine

Tamára L. Box, PhD, FAMIA
Deputy Executive Director, Analytics and Performance Integration, Office of Quality and Patient Safety, Veterans Health Administration

Shawn Forrest, MS
Digital Health Specialist, Food and Drug Administration, U.S. Department of Health and Human Services

Judy George, PhD
Lead, Quality Indicators Program, Division of Quality Measurement and Improvement, Center for Quality Improvement and Patient Safety, Agency for Healthcare Research and Quality

Michelle Schreiber, MD
Deputy Director of Center for Clinical Standards and Quality (CCSQ) and Director of the Quality, Measurement and Value-based Incentives Group (QMVIG), Centers for Medicare and Medicaid Services

ADVISORY GROUP MEMBERS

Helen Burstin, MD, MPH, MACP
Chief Executive Officer, Council of Medical Specialty Societies

Karen Cosby, MD, FACEP, CPPS
Medical Officer, Center for Quality Improvement and Patient Safety (CQuIPS), Agency for Healthcare Research and Quality

Patricia Dykes, PhD, MA, RN
Research Program Director, Center for Patient Safety, Research and Practice at Brigham Women's Hospital

Helen Haskell, MA
President, Mothers Against Medical Error

Catherine H. MacLean, MD, PhD
Chief Value Medical Officer and Senior Vice President, Hospital for Special Surgery

NATIONAL QUALITY FORUM STAFF

Elizabeth Drye, MD, SM
Chief Scientific Officer

Jenna Williams-Bader, MPH
Director

Heidi Bossley, MSN, MBA
Consultant

Deidra Smith-Fisher, MBA, PMP
Project Director

Emma Maclean
Senior Healthcare Analyst

Emily Bell, MPH
Analyst

Christine Craddock, MPH
Analyst

Appendix B: Methodology

GENERAL APPROACH

The process below describes how NQF and the TEP generated the strategies to advance trustworthy AI-enabled measures and recommendations for developing, selecting, and implementing these types of measures contained in this report:

1. Convened the multistakeholder TEP.
2. Gathered information relevant to the use of AI in quality measures by conducting a review of the literature, existing AI governance documents and frameworks, and consensus-based measure evaluation criteria, and holding several key informant interviews.
3. With the TEP, developed strategies to advance trustworthy AI-enabled measures and recommendations for the development, selection, and implementation of these types of quality measures in accountability programs.
4. Obtain public comment. (Current step)
5. Finalize strategies and recommendations. (Future step)

CONVENED THE MULTISTAKEHOLDER TEP

NQF seated a 17-member TEP representing diverse areas of interest and expertise, including liaisons from several federal agencies and NAM, following outreach to its membership and a broad public call for nominations. NQF included experts in the use of AI methods for healthcare quality measurement, purchasers and payers, health system providers using AI methods, patient advocates, experts in clinical informatics and health information technology, and experts in health equity related to the use of AI methods. NQF also consulted with a five-person advisory group composed of national leaders with different perspectives to guide the project.

GATHERED INFORMATION RELEVANT TO THE USE OF AI IN QUALITY MEASURES

To gather information relevant to the use of AI in quality measures, NQF:

- reviewed literature related to “artificial intelligence” and “quality measures”;
- identified AI governance and frameworks for healthcare applications and used these to inform the content in this report (*Appendix C*);
- reviewed current consensus-based measure evaluation criteria in order to inform recommendations for developing, selecting, and implementing AI-enabled measures; and
- conducted key informant interviews with experts in AI and quality measurement who have significant experience developing measures and implementing AI methods in healthcare applications.

DEVELOPED STRATEGIES AND RECOMMENDATIONS FOR THE DEVELOPMENT, SELECTION, AND IMPLEMENTATION OF AI-ENABLED QUALITY MEASURES IN ACCOUNTABILITY PROGRAMS

NQF convened the AI TEP in a series of virtual and in-person meetings in order to inform the strategies for advancing trustworthy AI-enabled measures and recommendations for the development, selection, and implementation of these types of measures contained in this report. NQF held three virtual

meetings with the TEP and identified an initial framework for the use of AI in quality measures. NQF further refined the initial framework into the six strategies contained in this report during a series of small group meetings with TEP members. With feedback from the TEP during an in-person meeting, and three subsequent web meetings, NQF synthesized the five-step process for developing, selecting, and implementing AI-enabled quality measures, using the strategies to ground the recommendations included in each step.

OBTAINING PUBLIC COMMENT AND FINALIZING STRATEGIES AND RECOMMENDATIONS

To gain additional feedback and further refine the draft strategies and recommendations, NQF is publishing the draft report for a three-week public comment period. Following the public comment period, NQF will summarize public comment for the TEP's review, finalize recommendations with the TEP, and publish the report.

Appendix C: Existing AI Governance and Frameworks

To understand the current state of AI governance and frameworks and to identify themes across frameworks, NQF reviewed several national frameworks and governance documents. NQF's key learnings from the documents that most influenced discussions with the TEP are listed below, organized by publishing entity.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

AI Risk Management Framework (2023)[31]

This framework, developed through a consensus-driven, transparent, and collaborative process between private and public sections, is a voluntary guidance document. It aims to improve the trustworthiness of AI systems; help organizations identify, assess, manage, and monitor AI risks; and support responsible and ethical AI development and use. The framework outlines characteristics of trustworthy AI systems, including: "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed." This framework underpins many other governance and guidance documents, such as the intervention risk management requirements for predictive decision support interventions in the ASTP HTI-1 Final Rule and CHAI's Blueprint for Trustworthy AI Implementation Guidance and Assurance for Health.

ASSISTANT SECRETARY FOR TECHNOLOGY POLICY

HTI-1 Final Rule (2023)[12]

In this rule, the Assistant Secretary for Technology Policy (ASTP, formerly the Office of the National Coordinator for Health IT [ONC]) describes requirements for health IT developers of decision support interventions, including evidence-based and predictive decision support interventions. The rule promotes transparency around AI algorithms used in decision support interventions by requiring developers to provide information about the:

- intervention details and outputs;
- intervention purpose;
- cautioned out-of-scope use of the intervention;
- intervention development details and input features;
- process used to ensure fairness in development of the intervention;
- external validation process;
- quantitative measures of performance;
- ongoing maintenance of intervention implementation and use; and
- update and continued validation or fairness assessment schedule

The requirements in this rule guided discussions with the TEP, emphasizing the importance of transparency regarding AI algorithms and identifying key aspects of the algorithm, its development and testing, and the data used to train and validate the algorithm that developers should provide.

U.S. FOOD AND DRUG ADMINISTRATION

Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) (2024)[13]

In this framework, the U.S. Food and Drug Administration (FDA) presents the concepts of "locked" and "adaptive" algorithms. Locked algorithms, which have traditionally been cleared or approved by the FDA, provide the same result each time the same input is applied to them and do not change with use. Changes in these types of algorithms "likely require FDA premarket review beyond the original market authorization." Changes to these algorithms are somewhat analogous to the field of quality measurement, in which measures that have changes to the measured outcome or process, population, data sources, setting of care, or level of analysis undergo review by consensus-based entities to confirm the measure is still scientifically sound.

The FDA notes how AI/ML prompts the need for a revised type of review because these algorithms may adapt over time as they continuously learn from real-world experience. The FDA's proposed approach to these adaptive algorithms is a "total product lifecycle (TPLC) regulatory approach that facilitates a rapid cycle of product improvement and allows these devices to continually improve while providing effective safeguards." Such an approach relies on a "predetermined change control plan," which describes the types of anticipated modifications, based on the retraining and model update strategy, and the associated methodology being used to implement those changes in a controlled manner that manages risks to patients. The TEP mentioned that a similar approach, using something like a predetermined change control plan (PCCP), could be applied to quality measures.

U.S. FOOD AND DRUG ADMINISTRATION, HEALTH CANADA, AND THE UNITED KINGDOM'S MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY

Good Machine Learning Practice for Medical Device Development: Guiding Principles (2021)[34]

Several of the guiding principles outlined in this document informed NQF's initial framing of TEP discussions about principles and recommendations. The principles highlight the following points:

- "Clinical study participants and data sets [or in the case of quality measures, patients included in the development data set for an AI-derived component] should be representative of the intended patient population."
- "Training data sets are independent of test sets."
- "Users are provided clear, essential information."
- "Deployed models are monitored for performance and re-training risks are managed."

U.S. FOOD AND DRUG ADMINISTRATION

Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations Draft Guidance for Industry and FDA Staff (2025)[35]

In this draft guidance, the FDA proposes a regulatory approach for AI-enabled device software functions (AI-DSFs), emphasizing a TPLC approach. The guidance supports the use of safe, effective, and equitable AI-DSFs by providing detailed recommendations for marketing submissions and lifecycle management. The TPLC approach encourages early consideration of transparency and bias mitigation, including performance evaluation across demographic subgroups and clinical settings. The guidance promotes the use of PCCPs to manage adaptive algorithms, enabling safe and effective updates to models.

Additionally, this guidance outlines documentation expectations across submission components, including detailed information on device description, risk assessment, model development, validation, performance monitoring, and public transparency. The FDA offers a summary template to communicate model characteristics, performance, and limitations.

NATIONAL ACADEMY OF MEDICINE

An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action (2025)[36]

The document encompasses a series of principles and commitments designed to guide the development and deployment of AI in the healthcare sector. These guidelines are intended for broad application across various stakeholders involved throughout the AI lifecycle. The principles comprise 10 key elements highlighting responsible AI development, use and ongoing monitoring, while providing touchpoints around which AI governance is to be shaped, tested, validated, and improved as technology advances. Several principles (provided below verbatim) are particularly relevant for quality measures, including:

- Engaged: Understanding, expressing, and prioritizing the needs, preferences, goals of people, and the related implications throughout the AI life cycle
- Equitable: Application accompanied by proof of appropriate steps to ensure fair and unbiased development and access to AI-associated benefits and risk mitigation measures
- Accessible: Ensuring that seamless stakeholder access and engagement is a core feature of each phase of the AI life cycle and governance
- Transparent: Provision of open, accessible, and understandable information on component AI elements, performance, and their associated outcomes
- Accountable: Identifiable and measurable actions taken in the development and use of AI, with clear documentation of benefits and clear controls and accountability for potentially adverse consequences
- Adaptive: Assurance that the accountability framework will deliver ongoing information on the results of AI application, for use as required for continuous learning and improvement in health, health care, biomedical science, and ultimately, the human condition

Additional principles apply to quality measures with some adjustments in definition

- "Safe: Attendance to and continuous vigilance and controls for potentially harmful consequences from the application of AI in health and medicine for individuals and population groups"
    - Currently, quality measures are applied retroactively to a data set to produce measure results and are therefore not used to determine treatment decisions for patients. In this way, they are "safe." However, as clinical decision support tied to measures influences clinical decisions and as AI and digital measures support real-time measurement, there may be the potential for AI-enabled measures to have an impact on patient care. In this case, it will be critical for the measures to be safe.

- "Effective: Application proven to achieve the intended improvement in personal health and the human condition, in the context of established ethical principles"

  - Quality measures do not directly impact patient health; they indirectly lead to health improvement through quality improvement interventions. Therefore, the definition of "effective" provided in the principles may not be applicable to measures. However, an AI-enabled measure should achieve its intended purpose of measuring what it is intended to measure and lead to improvements in healthcare delivery, experience, or outcomes, as defined by the measure.
- "Efficient: Development and use of AI that results in reductions in resources to achieve improved health outcomes without concomitant adverse impacts on the natural environment."
  - As stated above, quality measures do not directly result in better health outcomes. However, there are costs associated with developing and implementing measures, and measure developers are encouraged to develop a business case for each measure, which "predicts measure performance and the impact it will have on health and financial outcomes."[44] For quality measures, efficiency may involve weighing the benefits of using AI in the measure (e.g., to measure something previously unmeasurable or reducing reporting burden) to the costs (including resources needed) to implement the measure.

CONSUMER TECHNOLOGY ASSOCIATION

AI in Health Care: Practices for Identifying and Managing Bias (2023)[37]

This standard outlines best practices and guidance for identifying and minimizing bias in AI applications used in healthcare. It is designed to guide developers, healthcare providers, regulators, and other key stakeholders in recognizing and mitigating various forms of bias that can compromise AI applications. The standard categorizes various types of bias that can affect AI systems and identifies the stages in the AI lifecycle where these biases can be introduced, including the data collection and labeling phase. To mitigate these risks, the standard recommends a set of best practices known as "Good Data Management Practices," which include promoting transparency (i.e., clearly documenting the data sets and algorithms used), encouraging diversity throughout the development lifecycle (i.e., including a variety of perspectives within the development team), using representative data (i.e., ensuring datasets include key demographic elements) screening and auditing for bias (i.e., defining the algorithm's purpose before development and execution), and retraining algorithms (i.e., evaluating and applying strategies to minimize bias including training with new or updated data). While the document synthesizes current best practices, it acknowledges the rapidly evolving landscape of AI in healthcare and advises users to stay informed about applicable federal, state, and local regulations.

The Use of AI in Health Care: Trustworthiness (2020)[38]

This standard outlines the core requirements for AI solutions in healthcare to be deemed trustworthy. The standard identifies three key dimensions of how trust is created and maintained: human trust, which emphasizes usability and the relationship between users and developers; technical trust, which confirms the AI systems are designed and trained to perform as expected; and regulatory trust, which involves adherence to laws and regulations designed to prevent harm to end users. The standard emphasizes the importance of trust from the end user perspective, including physicians, consumers, caregivers, public health officials, medical societies, and regulators.

COALITION FOR HEALTH AI

Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare (2023)[32]

The blueprint supports the use of trustworthy AI in healthcare by identifying and proposing solutions to issues that must be addressed. The consensus-based recommendations, informed by a coalition of experts from healthcare systems, academia, government, and industry are designed to enhance trustworthiness and promote responsible adoption of AI technologies within the healthcare sector. The report builds from the National Institute of Standards and Technology (NIST) AI risk management framework, providing definitions of key terms related to AI and how they apply in a healthcare context, as well as describing issues that will impact the ability to build and implement trustworthy AI. The document defines terms such as "valid," "reliable," "reproducibility," "monitoring," "transparency," "bias," and "fairness," some of which have similar definitions in the quality measurement context and some which differ in definition when applied to quality measures.[31]

Responsible AI Guide (2024)[57]

The detailed playbook offers best practices for the development and implementation of trustworthy AI and is targeted at a broad audience, including those selecting, developing, and implementing AI technologies in the delivery of patient care and related health system processes. The guide outlines and is organized around a six-stage health AI lifecycle:

- Define problem and plan
- Design the AI system
- Engineer the AI solution
- Assess
- Pilot
- Deploy and monitor

The guide outlines considerations for each step which are further organized around five "principle-based themes":

- Usefulness, usability, and efficacy
- Fairness
- Safety and reliability
- Transparency, intelligibility, and accountability
- Security and privacy

Appendix D: Completed Quality Measure AI Model Summary Label Example

Quality Measure AI Model Summary Label Example for AI-Derived Component in Quality Measure

AI-Derived Component Information

- **Name of the AI-derived component:** Adenoma Detection Rate (ADR) NLP Extraction System.
- **Name of the developer of the component (may or may not be the measure developer):** NLP Vendor.
- **Version of the component used in the measure (i.e., model/software release version):** NLP Vendor Monitor ADR 3.20.
- **Date when the component was created (or last updated):** 6-21-23.

Description

- **Intended users (e.g., healthcare providers, health plans, caregivers, patients):** Healthcare quality analysts, gastroenterologists, clinical operations, and reporting teams.
- **Intended use: The general purpose of the component or its function. This includes descriptions of how the component is used in the quality measure, the target patient population for which the component is intended, and the intended care setting(s) in which the component is used (e.g., hospital, ambulatory care):** Automated extraction and classification of clinical concepts (e.g., problems, procedures, medications) from unstructured colonoscopy and pathology reports to compute ADR quality metrics to support quality measurement and clinical reporting. This pipeline targets general inpatient and outpatient care settings and aims to support populations in medical-surgical, oncology, and primary care domains.
- **Instructions for use: Directions and recommendations for optimal use of the component in the measure by the measured entity:** Upload clinical documents into the system. The conditional random field (CRF) based model extracts clinical entities from text, and the classification model assigns document- or entity-level labels (e.g., clinical relevance, assertion status). Outputs are reviewed through a validation interface or fed into downstream quality reporting pipelines.
- **Rationale**: **The rationale for using the component in the quality measure, including a description of the clinical or quality concept that it attempts to capture, why AI should be used to capture the concept rather than other methods (e.g., administrative data, EHR data), and how the resulting definitions, associated coding and terms, variables, and other inputs represent the clinical concept:** The ADR metric is a key quality measure in gastrointestinal (GI) care, indicating the percentage of screening colonoscopies in which at least one adenoma is detected. Accurate and timely calculation of ADR is critical for assessing provider performance, meeting CMS MIPS-343 reporting requirements, and improving patient outcomes through early polyp detection (https://www.nejm.org/doi/full/10.1056/NEJMoa1309086). Traditional methods to compute ADR rely on manual abstraction of colonoscopy and pathology reports or on administrative data, which is time-consuming, costly, and prone to error. Administrative data sources often lack the clinical specificity required to distinguish screening vs. diagnostic colonoscopies or to identify adenoma subtypes. The combined use of CRF-based named entity recognition (NER) and a downstream classification model enables precise, scalable extraction of clinically relevant entities and their contextual classifications (e.g., confirmed vs. negated findings). This automated approach reduces

burden on clinical abstractors, improves data consistency, and enhances the timeliness and quality of extracted clinical information.

- **Type of algorithm/model, including whether the component is predictive or generative, and a description of how it interacts with other systems (e.g., EHRs, integrated platforms, patient-generated information):** The pipeline includes: (1) A CRF-based extraction model for identifying clinical entity spans (problems, medications, procedures) and (2) A classification model to categorize each entity into relevant clinical categories. The system is non-generative and predictive, interacting with clinical data repositories and feeding structured outputs into quality reporting, clinical analytics platforms, and registry submissions. Its primary data sources are colonoscopy procedure reports and pathology reports containing histopathologic diagnoses. These reports are rich in clinical detail but are traditionally stored as unstructured free text within the electronic health record. In addition to the main report narratives, the system leverages supporting metadata such as report dates and authoring provider information, ensuring accurate metric attribution and time alignment. The system is also grounded in a comprehensive clinical ontology, enabling it to recognize and extract domain-specific terminology such as adenoma subtypes and procedure indications that are critical to quality metric computation. Together, these diverse data inputs support a scalable, automated approach to calculating Adenoma Detection Rate and related quality metrics.
- **Inputs: A description of the data source(s) used as inputs by the component, including the source of data that are necessary as input into the component and the types of data used (e.g., EHR, imaging):** Unstructured clinical documentation (e.g., colonoscopy reports, pathology reports) from supported electronic health record (EHR) systems. Metadata such as encounter date, provider specialty, and care setting are optionally used to improve classification accuracy.
- **Outputs: A description of the outputs of the component, including the type and value, and whether the output is a prediction, classification, evaluation, analysis, or another form:** The ADR NLP system generates structured outputs that enable automated quality measurement in gastrointestinal care. By extracting key clinical variables—such as colonoscopy type, diagnosis details, polyp size, and procedure completeness—the system transforms unstructured clinical text into actionable data. These variables are then used to automatically calculate quality metrics, including the CMS MIPS 343 Adenoma Detection Rate, overall adenoma detection, and advanced adenoma rates. The outputs of the system include both individual classifications (e.g., whether an adenoma was detected in a screening colonoscopy) and aggregated metric rates that summarize performance across patient populations and individual providers performing the procedure. This allows healthcare teams to classify cases accurately, compute quality performance indicators, and evaluate outcomes for quality improvement. The system supports both retrospective quality reviews and prospective monitoring, reducing the need for manual chart abstraction and enhancing the timeliness of quality reporting.

Development and Testing

- **Characterization of data used to develop and test the component (these data sets should be separate):**
    - **Data sources (e.g., health system data, public or proprietary databases) including details on any devices used to collect data**: Academic Medical Center Electronic Health Record

system (Epic). Documents were retrieved using proprietary document repository search engine using procedure and diagnosis keywords.

- **Data types used (e.g., structured numerical data, structured categorical data, unstructured text, images, time-series data, or a combination):** Primary data type is unstructured text in the form of free-text clinical reports. Secondary data includes structured metadata, document timestamps, report authors, and procedure metadata.
- **Pre-processing applied to data before developing the component**: Deduplication of report pairs, de-identification, sentence segmentation, section segmentation, and tokenization. Ground truth labeled using ontology-guided annotation and domain expert review.
- **Relevant details including:**
    - **Unit of analysis:** Patient-level colonoscopy and pathology report pairs.
    - **Number of patients/records/data points:** 2,500 clinical notes, split across training, validation, and test sets
    - **How the developer sampled the data, if applicable:** Targeted sequential sampling to achieve desired distribution of diagnoses and procedures.
    - **A description of the data sources that were available in the data set but not included and why the developer did not include them:** Colonoscopy reports generated outside of the designated source system were excluded due to lack of consistent formatting and initial scope
    - **Characteristics of patients included in the data set:** Adult population 50 years and older undergoing colonoscopy.
    - **Characteristics of patients excluded from the data set:** Pediatric patients, reports outside of supported EMR.
- **A description of subpopulation characteristics (e.g., the percentage of subgroups captured by the component) and an assessment of whether the data can be considered representative of the overall intended population:** The dataset used to develop and test the clinical entity extraction and classification pipeline primarily represents an adult patient population receiving care at a single urban academic medical center encompassing both rural and urban populations across 22 individual GI laboratories. Most patients are between 50 and 75 years old, with a balanced distribution of male (49%) and female (50%) patients, and a small proportion identifying as nonbinary or other genders (1%). Racial and ethnic demographics reflective of the center's diverse population: approximately 45% of patients identify as White, 25% as Black or African American, 15% as Hispanic or Latino, 10% as Asian, and 5% as other or mixed race. The patient cohort includes a range of insurance types (commercial, Medicare, Medicaid, and uninsured), and roughly 20% reside in rural ZIP codes based on RUCA classifications. The main limitation of the population is that it represents a single academic center within one tri-state region. Additionally, rare diseases and highly specialized procedures are less prevalent in the training data. As a result, the data is considered generally representative of adult tertiary care environments but may require further validation for use in other care settings or populations.
- **Characteristics of healthcare entities included in the data set:** Large urban academic medical center in the Midwest. The dataset included clinical documents from a single large urban academic medical center in the Midwest, covering inpatient, outpatient, and procedural care settings.

- **Characteristics of healthcare entities excluded from the data set:** Data from the affiliated community clinics, rural hospitals, and long-term care facilities were excluded due to differences in documentation practices, inconsistent availability of structured clinical narratives, and the initial project scope.
- **A description of the process for developing the component**: The development of the clinical entity extraction and classification pipeline followed a structured, iterative process combining domain expertise and machine learning best practices. Initially, a corpus of clinical documents was collected from an academic medical center's electronic health record system. These documents were pre-processed through de-identification, sentence and section segmentation, and tokenization. Clinical entities were annotated by domain experts using a predefined ontology covering problems, procedures, and medications, along with context attributes such as assertion status. A CRF model was trained on the annotated dataset to perform NER, identifying relevant spans of clinical text. Ontology terms (e.g. hyperplastic polyps, serrated sessile polyps, villous adenomas) and classification categories (e.g. those meeting the definition for inclusion in the metric) were iteratively refined based on error analysis and expert review. The development process incorporated separate training, tuning, and testing phases, with each phase using distinct subsets of the data to prevent data leakage. Final model performance was evaluated on a held-out test set of documents that were not seen during training or tuning. The pipeline was optimized for both accuracy and generalizability within the target clinical environment. Ongoing feedback from clinical informatics teams informed additional refinements prior to production deployment.

- **Description of how missing data and/or a limited data set may impact performance of the component:** The ADR NLP system relies on complete colonoscopy and pathology reports. Missing or incomplete documentation will result in failure to extract key variables, preventing metric calculation. Less common diagnoses and rare clinical findings may be underrepresented in the data set, learning to lower extraction accuracy.
- **Limitations of the data sets used for development and testing, including if the developer needed to normalize or translate the data:** All data originated from a single large academic institution, limiting the variability in report style and terminology seen across other health systems. Basic text normalization was applied during NLP preprocessing.

Performance

- **A description of the process used for testing the performance of the AI-derived component and a description of the types of tests used**: The performance of the ADR NLP component was evaluated through a structured, multi-phase process designed to simulate real-world data extraction and quality metric computation in a healthcare environment. A total of 2,500 cases containing colonoscopy and pathology report pairs were collected from the EHR system using target keyword queries. Cases were sequentially sampled to achieve a representative distribution of diagnoses and procedure types, ensuring clinically meaningful diversity.
- **A summary of the performance results:** The ADR NLP component demonstrated high overall performance in extracting clinically relevant variables from colonoscopy and pathology reports. Key extraction tasks, including identifying colonoscopy exam type, adenoma subtypes, and polyp size,

achieved F1 scores ranging from 0.85 to 0.99, with particularly strong performance on variables critical for ADR metric computation. Additional evaluation metrics for extraction tasks include: Sensitivity (Recall): 0.87-0.98, Specificity: 0.89-0.99, Precision: 0.86-0.98. The system showed excellent accuracy for screening colonoscopy classification and adenoma detection. Sentence boundary detection accounted for a smaller proportion of errors. Performance on rare or complex findings was slightly lower, with F1 scores in the 0.78-0.84 range. However, these had minimal impact on the overall ADR metric calculations due to their low frequency and limited effect on denominator counts.

- **Stratification of the testing results by patient characteristics: *Clinical Findings*:** 35% Adenomas, 15% Serrated Adenomas, 15% Advanced Adenomas, 15% Cancers, and 20% Non-diagnostic cases. The target distribution for test set procedure was 75% Screening and 25% Non-screening. ***Age***: 50–64 years: 52%, 65–75 years: 30%, Over 75 years: 18%. ***Gender***: Male: 49%, Female: 50%, Nonbinary / Other: 1%. ***Race and Ethnicity***: White: 45%, Black or African American: 25%, Hispanic or Latino: 15%, Asian: 10%, Other or Mixed Race: 5%. ***Geography***: Urban ZIP codes: 80%, Rural ZIP codes: 20%. ***Insurance Status***: Medicare: 60%, Commercial: 25%, Medicaid: 10%, Uninsured: 5%.
- **Links to published evidence describing development and/or testing of the AI-derived component**. None

Risk Management

- **Potential risks associated with the component, the data, and the outputs (e.g., bias risks, information gaps):** Reduced accuracy on reports from unsupported EHR systems. Possible misclassification of colonoscopy type when multiple indications are present. Possible underrepresentation of rare diagnostic entities in training data. Reliance on domain-specific ontologies that may miss novel synonyms or phrases.
- **Interactions, deployment, and updates. When appropriate, provide:**
    - **Resources required to implement the component, including computational resources, IT infrastructure, staffing expertise and numbers, and whether there is a cost to license the component:** Implementing the clinical entity extraction and classification pipeline requires a combination of computational resources, IT infrastructure, and specialized staffing. The system runs efficiently on a moderate compute environment, typically using a 4 to 8 vCPU application server for hosting the API and performing preprocessing tasks, along with either a mid-tier GPU or a CPU-only setup with 16 to 32 GB of RAM for model inference. Approximately 1 TB of storage is recommended to support document archiving, log storage, and temporary processing. The solution integrates with existing clinical document repositories or EHRs, operating within a secure, HIPAA-compliant environment, either on-premises or in a healthcare-compliant cloud platform such as AWS, Azure, or Google Cloud. Containerization technologies like Docker or Kubernetes are used for deployment, supported by CI/CD pipelines and standard monitoring tools for service reliability. The implementation team typically includes one full-time machine learning or NLP engineer to maintain the models and pipelines, a clinical informaticist or data scientist (half to full-time) to refine the clinical ontology and validate extracted entities, and a part-time DevOps or cloud engineer to manage deployment, scaling, and infrastructure health. A clinical domain expert provides periodic guidance on terminology and reviews the system's performance

against clinical standards. Additional support from a project manager may be needed to coordinate implementation efforts and maintain communication with clinical and operational stakeholders. Depending on the organization's needs, optional roles such as QA analysts or abstractors may assist with testing and annotation refinement.

- **Details regarding how the component is deployed and updated, including:**
  - **How to conduct local site-specific acceptance testing or validation:** After model training and tuning are complete, the ADR NLP component undergoes site-specific acceptance testing to ensure accuracy and reliability within the local data environment. This process begins with deployment to a staging environment, where the production model is used to process a fresh sample of clinical reports. These reports are distinct from the original training and test sets, ensuring an unbiased validation. Following successful staging validation, the model is deployed to production and subjected to a smoke test using newly ingested production data. This final check ensures that the system functions correctly in the live environment before full production signoff.
  - **Ongoing performance monitoring and maintenance:** After deployment, ongoing performance monitoring is conducted through continuous review of system-identified errors. All false positives flagged by human abstractors are reviewed and analyzed by a domain expert. Each false positive is assigned an internal error code to classify the root cause. These findings are regularly evaluated to identify opportunities for improvement. Model updates and refinements are prioritized for inclusion in subsequent model releases, enabling the system to adapt to evolving clinical language and documentation patterns. This feedback loop ensures sustained system accuracy and responsiveness to real-world use.
  - **Transparent reporting of successes and failures:** To promote transparency and trust, model performance is reviewed bi-yearly with key stakeholders and customers. These sessions provide a forum to discuss both successes and areas for improvement. During each review, the team shares field precision metrics, summarized identified false positives and false negatives, and reports on the root causes of recent errors. Proposed changes to the model and any associated risks are discussed collaboratively. All model updates are reviewed, assessed, and approved by clinical, technical, and operational stakeholders before inclusion in future releases. This continuous feedback and governance process ensures that model improvements align with clinical goals and operational priorities.
  - **Change management strategies:** The clinical entity extraction and classification pipeline is deployed using a controlled, phased release process supported by standard change management practices. All code and model updates follow a formal CI/CD workflow, where changes are version-controlled, peer-reviewed, and automatically tested against regression suites before deployment.
  - **Proactive approaches to address vulnerabilities:** Vulnerability management of service included automated scanning and detection. Performing standard maintenance including scanning and patching.

  - **Communication to parties of as-needed information:** Communication for ADR NLP is decided by the ADR communication plan which details the audience, cadence, and content of communication regarding the service.
  - **Software quality (specify, standards and regulatory compliance issues, intellectual property issues, risk management and safeguards used, other):** HIPAA and HITRUST compliant and certified
- **Known risks, biases, or failure modes**: The clinical entity extraction and classification pipeline is subject to several potential risks and failure modes. A key risk is reduced accuracy when processing clinical notes that differ substantially from those used during development, such as notes from other institutions with different documentation styles, section headers, or terminology. Additionally, the model may struggle with rare clinical concepts or unusual phrasing not well represented in the training data, potentially leading to missed extractions or incorrect classifications. Misclassification of clinical assertions (e.g., incorrectly identifying a condition as present rather than negated) represents another common failure mode. The system also depends on consistent document formatting and sectioning; poorly structured or fragmented documents may degrade extraction performance. Biases may arise from the limited diversity of the development dataset, which was drawn from a single academic medical center. This introduces the potential for demographic and clinical practice biases, as rural, pediatric, or underrepresented patient populations and specialties may be less well represented in the training data, leading to lower performance on these subgroups. Furthermore, entity classification categories are based on domain expert input from a single institution, which may reflect local clinical practices and not generalize universally.
- **Bias mitigation approaches used during development and testing of the component.** To mitigate risks, the development process included stratified sampling to ensure a balanced representation of common clinical specialties and diagnoses. The annotation guidelines were reviewed iteratively to reduce conceptual bias and clarify ambiguous cases. The classification model was evaluated for performance consistency across key subgroups (e.g., by care setting and diagnosis category). During testing, the system's errors were analyzed for patterns suggesting bias, such as disproportionately low recall in certain specialties, prompting additional training data collection where feasible. The team also prioritized transparency by documenting known data limitations and failure modes.
- **Known circumstances where the input for the component will not align with the data used in development and validation:** The pipeline is optimized for clinical documentation generated at an academic medical center and may encounter alignment issues when applied to documents from other healthcare settings. The system may not align well with transcribed clinical dictations, scanned documents converted by OCR, or documents written in languages other than English. These situations may result in degraded extraction accuracy and increased classification errors, requiring further local validation and tuning before deployment in those settings.
- **Ethical or clinical implications that may arise from component misclassification:** Clinical risk level is Low; results are intended for quality metric reporting, not for direct clinical decision-making.