Standards and Quality for Artificial Intelligence Tools

August 2023

AI

VIEWPOINTS

Exploring quality assurance for AI diagnostic decision support technologies

As artificial intelligence (AI) technologies continue to proliferate across business, government, and society, they are no longer a futuristic concept but a day-to-day reality with which leaders must contend. Spurred by the development of models such as ChatGPT, AI is now at the conversational forefront of many industries, including healthcare, where it has the potential to decrease workforce burnout, increase efficiency, improve patient outcomes, and reduce costs.[1] However, the speed with which healthcare vendors and systems are developing and deploying AI, coupled with the lack of standards for building algorithms and validating performance, has raised questions about the quality and quality assurance (QA) of these technologies and the potential effects on patient outcomes.

Machine-learning diagnostic decision support (ML-DDS) tools are a subset of AI technologies that are currently being utilized in a broad range of clinical settings, from assisting in breast cancer detection to assessing a patient's risk for Alzheimer's disease. Diverse organizations are advancing the creation and use of ML-DDS in the market. Digital health start-ups with AI decision support raised $1.2 billion in 2022, an amount higher than any other category of AI in healthcare.[2] Large technology companies such as Microsoft, Philips, and Google are developing ML-DDS products for healthcare providers and systems. Health systems themselves are also building ML-DDS tools internally as in-house tools offer several benefits for health systems: the datasets employed are relevant to specific health systems' respective populations; there can be more seamless integration with existing IT infrastructure in comparison to third-party offerings; and there is less regulatory scrutiny as the US Food and Drug Administration (FDA) has not historically exercised oversight over hospital-developed solutions. These factors have cultivated a preference for adequately resourced health systems to build, not buy, ML-DDS technologies.[3]

Recent events have raised concerns about how to build and integrate ML-DDS products into patient care. For example, Epic's sepsis model was used by hundreds of hospitals to help diagnose sepsis, but real-world performance revealed the technology to be substandard in some settings, leading to diagnostic errors and potential patient harm.[4] Furthermore, the lack of diversity in datasets used to develop ML-DDS has garnered significant attention, with particular apprehension around how AI bias can lead to diagnostic inaccuracies for underrepresented groups.[5]

The FDA has recently signaled an intent to regulate all software with diagnostic or treatment-related outputs, thereby setting baseline requirements for safety and performance for all ML-

DDS solutions in the future. However, there remains uncertainty around how such regulations would be implemented.[6]

Against this backdrop, Tapestry Networks, with support from the Gordon and Betty Moore Foundation, held a series of one-on-one and small-group conversations with key stakeholders, including AI developers and vendors, payers, and academic and community health systems. These participants offered their views on the following:

- **What are the opportunities and challenges in advancing the safe, effective, and fair adoption of ML-DDS?** (page 2)
- **How can potential demonstration projects progress the quality of healthcare AI?** (page 6)

This *ViewPoints* provides a synthesis of those conversations and integrates additional research where relevant. *For a full list of participants, please see Appendix 1, on page 17.*

Opportunities and challenges in advancing the safe, effective, and fair adoption of ML-DDS

Stakeholders saw great promise in AI, but nearly all highlighted critical gaps in standardization and QA. In discussing the opportunities and challenges that healthcare AI presents, many saw a need for the establishment of a shared set of guiding principles and good practices to improve the quality of such technologies. However, further discussion revealed the challenges involved in finding neutral, transparent, and credible entities to move this vision forward.

Stakeholders universally affirmed the need for quality standards

Stakeholders are excited about AI, but to ensure that the benefits of such technologies are realized in healthcare, all stakeholders agreed on the need for standards and QA, particularly for tools such as ML-DDS, which can directly affect patient care and outcomes. Discussions on gaps in standards and quality yielded the following insights:

- **Many are optimistic about ML-DDS's benefits in healthcare.** One self-insured employer shared how ML-DDS may positively impact their patient population with respect to reducing diagnostic error: *"Last year, we saw diagnostic error rates of over 30% for our patients, so I'm excited about AI, especially as studies are showing improved diagnosis rates in areas like radiology and pathology. We need some better way to help our patients get the correct diagnoses and on to the right treatment pathways."*

> *"There's as much bad science as there is good science out there. I do think there is a need for quality regulation."*
>
> —Industry representative

- **Stakeholders also acknowledged the pitfalls of AI, including the potential for bias, and expressed concern around the lack of standards for development and evaluation of ML-DDS.** *"There's no governance on quality across institutions right now—the field is moving quickly and there's not a lot of energy put into thoughtful development and deployment of these technologies.*

It may only be a matter of time until there's a big overstep or care issue that leads to a national news story," one stakeholder commented. A payer echoed similar sentiments, underscoring the need for algorithm validation and oversight in light of the potential for bias, among other issues: *"At the moment, there's no entity that looks at how an algorithm is built, so if there's junk data in, then it'll just be junk out. There are also no standards to look at performance over time, and we know algorithms degrade. I wouldn't want any patient diagnoses to be informed by something that has a lot of error, bias, and performance reliability."*

- **Stakeholders lack clarity around how quality standards for diagnostic AI should be set.** Some noted the need for clarity on how standards should compare the diagnostic accuracy of AI technologies with that of clinicians. One said, *"We have been deploying AI across hundreds of clinical practices, and the quality metrics we've been asked to meet are substantially higher than the quality we see in physicians. The Epic algorithm has been in the news a lot for its low detection rates, but it performs better than a lot of doctors. We're very quick to judge AI quality, but in many areas, we are not measuring physician quality. It's a real paradox."*

There is uncertainty around who should lead QA for diagnostic AI

Despite agreement on the need for quality standards, stakeholders had mixed views on which institution or entity should lead the development and evaluation of such standards. Many noted that while, in principle, an independent, transparent, and trusted organization with appropriate expertise should steer a sector-wide quality initiative, gaps in resourcing and current market incentives may mean that stakeholders with conflicting interests assume leadership roles.

Interview participants discussed various considerations around the optimal profile of quality leaders for healthcare AI:

- **FDA.** While the regulatory agency is a logical fit to lead QA for diagnostic AI, stakeholders believe the agency has been *"under-resourced and overwhelmed for many years."* Even as the FDA moves to expand oversight of AI technologies, some stakeholders—payers in particular—lacked confidence in the agency's assessments for AI and digital technologies more broadly. One said, *"The agency's standards for approval or clearance [for digital technologies] are now so low that they're almost useless. We still do a lot of work when we review FDA-approved devices,"* highlighting the difference in standards used by the federal agency for initial clearance and the standards set by stakeholders for real-world implementation. Others shared apprehensions around the regulatory agency's ability to keep pace with change: *"The rate of acceleration for AI-based technologies is now at warp speed, and it's hard to see how traditional regulatory approaches for quality can adapt themselves appropriately."*

- **Academic medical centers.** Some believe that academic medical centers (AMCs) could play a role in setting standards for diagnostic AI, given that many are spearheading research and development of ML-DDS. However, numerous stakeholders cautioned against this approach, with one clinician describing an instance where standards developed by AMCs have led to complexities during implementation: *"A new reporting standard in pathology has been instituted recently, but in community settings, the parameters can lead to patients being overtreated with chemotherapy. I've seen this type of situation a lot when academics try to impose their agenda outside of their systems."* In short, as AMCs represent a relatively small percentage of sites where patients receive care, widespread application of standards from such institutions can be a case of *"the minority trying to police the majority,"* in the words of one interviewee.
- **Professional societies.** Societies already set clinical guidelines and standards, and in some cases also provide accreditation for health services; therefore, some stakeholders noted that societies' oversight of AI may be a natural extension of their existing work. The American College of Radiologists and the Radiological Society of North America both played pivotal roles in progressing the use of Digital Imaging and Communications in Medicine (DICOM) standards, advancing the protocol for data exchange and illustrating the capability of professional societies to drive change and standardization across the sector.[7] At the same time, stakeholders said that professional societies, while ostensibly representing a broad swath of relevant clinicians, can in fact reflect narrow interests: *"I would say the [professional society] represents an academic point of view. Since only 20% of patients are seen in academic settings, in my opinion, the organization lacks perspective on different types of patients and sites of care."* Additionally, while payers and other stakeholders look to professional societies for standards on quality, many doubted societies' ability to be neutral evaluators of AI, with one payer noting that *"societies can be too self-serving, so guidelines from one society need to be considered with a grain of salt,"* and in the broader context of guidelines from other societies and organizations.
- **Payers.** Some stakeholders anticipate that payers and large self-insured employers will eventually take the lead in quality surveillance. Payers and employers have access to the claims and outcomes data necessary to evaluate quality and, given their national reach, can set and implement standards broadly. Additionally, as healthcare costs continue to rise and value-based payment models take hold, some participants reasoned that payers and self-insured employers have, in principle, the financial incentive to evaluate AI products: *"Ultimately, they are the ones paying and have to deliver a higher quality of care."* However, commercial payers reimburse very few diagnostic AI technologies today because of a lack of evidence of improved patient outcomes. As a result, many payers currently take a hands-off approach, relying instead on clinicians to evaluate the quality of ML-DDS. *"We care about quality, but just like we don't dictate the tools used for surgeries, we also do not want to dictate the software clinicians should be using,"* one payer said. One interviewee highlighted the challenge with this payer approach: *"The irony is that payers say they'll let*

> *providers decide, but providers are unable or unwilling to make the investments necessary to improve quality until they are reimbursed better. It's the same chicken-or-the-egg situation that we see in many other areas of healthcare."*

While there remains uncertainty around the optimal institution to lead on quality in AI in healthcare, many stakeholders acknowledged that a centralized standard-setting and evaluative body for AI and ML-DDS for the sector would be ideal. One payer said that *"there's a huge need for an independent assessment entity,"* while another noted the potential value of *"an informed third-party organization without financial interests at hand."* However, stakeholders also noted the inherent challenge in establishing such an institution, particularly in the current market: *"Everyone is trying to sell something with AI baked in right now. If they can gain traction without showing any differential in outcomes, why would companies care about demonstrating quality? There are no incentives for that."*

Key principles should guide diagnostic AI quality advancement

Despite the lack of a centralized body to lead on diagnostic AI quality in healthcare, stakeholders communicated the importance of *"taking small steps today, through individual projects, to progress the field."* Some even said that their interest and work in quality is a direct result of the current *"Wild West"* landscape: *"Either we wait until people start to demand standards and transparency for AI and do nothing now, or we push to get the ball rolling."*

Across all conversations, stakeholders described several principles that should underlie the advancement of quality that future leadership bodies, when eventually identified, will need to consider:

- **Multistakeholder collaboration.** Almost all participants highlighted the importance of cross-stakeholder conversations for any quality-related initiative in the ML-DDS space. One said, *"Sometimes standards are developed without the right people at the table. Payers should help determine what's implementable in the claims system. Industry should guide how quality can be driven through platforms. Scientists should provide the necessary expertise."* One clinician with AI expertise noted, *"There are so many academic papers on quality frameworks and models, but very few of them have been adopted because they were developed without input from those it would impact, and they fail to work in the real world."*
- **Scalability and applicability across sites of care.** Stakeholders noted that the patient care landscape is shifting rapidly. In radiology, for example, an increasing number of patients receive services in imaging facilities owned by radiology conglomerates and retailers, and not in hospital systems.[8] Similar shifts in sites of care are also occurring in other specialty areas: *"What people miss is that so much patient care is happening outside of health systems and in large private groups instead. Cardiology, musculoskeletal, and women's health are all establishing themselves outside of academic centers."* Given these changing market dynamics, there will be a need to consider how quality can be evaluated and assured across different care contexts.

- **Adaptability to technological change.** AI and ML-DDS tools may be relatively new to healthcare, but even as stakeholders grapple with quality issues for current software, the technology continues to advance rapidly. Generative AI, for example, is top of mind for many: *"Consider the possibilities—there may be a future when a clinician walks into a room and the AI has already summarized the three most important things to address with the patient."* Some believe it will be a long time before generative AI is utilized for diagnostic tasks, but for a future QA leader in this space, the pace of change will require that they continuously consider the future in parallel with standards development. *"We have to consider the foundational elements and develop an approach which can be used time and time again when considering all forms of new AI technology,"* one stakeholder explained.
- **Sustainability of initiatives.** Stakeholders acknowledged that diverse and often competing incentives can encourage, or fail to encourage, the adoption of AI today. In the pursuit of quality improvement, some highlighted the need for pragmatism and an appreciation for commercial sustainability to motivate multistakeholder participation and contribution. One stakeholder said, *"It's great to have those interested in quality and better patient care volunteer their time to improve the status quo, but to shape innovation, somehow everyone along the pipeline has to have some wins." Further discussion on the role of incentives and financial reimbursement can be found on page 13.*

Potential demonstration projects to progress the quality of diagnostic AI

With these general opportunities and challenges in mind, interviewees considered how specific projects could benefit the field broadly and respond to some of the quality gaps noted above. Of note, the proposed concepts focus on quality improvement and QA with a specific subset of stakeholders–health systems, clinicians, vendors, and AI developers–given their existing thought leadership on healthcare AI and engagement with philanthropic institutions, including the Moore Foundation. Additional levers to impact AI quality also exist, including payer and consumer-driven efforts, as well as the creation of new regulatory, national, and international policies. *For a full list of the initial concepts discussed, please see Appendix 2, on page 19.*

Evaluating potential demonstration projects

Based on stakeholder feedback on seven initial project concepts and broader market engagement, the Gordon and Betty Moore Foundation sought further multistakeholder input on three specific proposals. All three proposals aim to serve as demonstration projects, with the goal of generating real-world evidence on the concepts of transparency, validation, evaluation, and monitoring that may be used in future governance approaches to QA for healthcare AI. Two proposals center on health delivery settings—specifically, evaluating the AI readiness of health systems and developing an AI best-practices resource for decision-makers. A third proposal focuses on AI quality standards and evaluation of real-world performance within a specialty.

Tapestry facilitated multistakeholder group discussions on how each proposal could be refined and improved to enable greater impact; a summary of these conversations follows. *For a list of review criteria for these proposals, please see Appendix 3, on page 20.*

Proposal 1: AI maturity model for health systems

At a time when health systems are being inundated with AI technologies, stakeholders see the value of an AI maturity model for organizational assessment, as detailed below.

Overview of Proposal 1

To develop, implement, and monitor AI technologies effectively, health systems need to have the right governance, infrastructure, and data analytics systems in place. However, there is currently no standardized evaluation that health systems can use to benchmark their capabilities, though some organizations are working towards such measures.[9] In contrast, the Healthcare Information and Management Systems Society (HIMSS) has created maturity models to evaluate electronic medical record adoption and analytical capabilities: the Electronic Medical Record Adoption Model (EMRAM) and the Adoption Model for Analytics Maturity (AMAM), respectively. Both models are utilized globally and score health systems from stage 0 to stage 7, with stage 7 being the highest level of maturity. EMRAM considers factors such as availability of lab data, percentage of clinical documentation produced electronically, and point-of-care IT infrastructure, while AMAM analyzes governance, analytics life cycles, and impact on patient outcomes.[10]

A multistakeholder coalition is proposing the development of a maturity model to evaluate the AI readiness of health systems to support AI development and deployment. **The goal is to enable health systems to identify strengths and weaknesses internally—though the maturity score may also be used to compare one system against another—and to progress the quality of AI by facilitating recognition of where readiness improvements need to be made.** While the model will be built at an AMC, it will also be tested in community health systems and modified accordingly ahead of large-scale deployment.

Many underscored the benefits of such a project, noting the importance of a model specifically for AI. One health system leader said, *"We are approached by AI vendors all across the organization. Leaders are excited by the new possibilities with AI, but they don't understand what is required for those possibilities to be realized. Our organization has a lot of work to do* *internally, and a model could help us understand exactly where we need to grow."* Another agreed that the model could *"provide health systems with an organized road map to work with"* and also allow *"a level of public transparency and trust with new technologies."* Building trust in an environment of *"tremendous skepticism and disbelief of the healhcare system,"* as one clinician opined, may be critical to the widespread adoption of AI and ML-DDS.

However, while all agreed on the need for an assessment methodology for health systems, stakeholders raised several points for consideration:

- **AI maturity varies across health system specialties, sites, and departments.** Given the breadth of specialties across many different sites, stakeholders were uncertain if one assessment methodology would be sufficient. One said, *"You can look at an institution's EMR [electronic medical record] maturity because it's all one system, but with AI, one department could be mature and ready while another might have no idea what AI is."* Additionally, AI technologies have developed at varying rates within specialties such that there are *"inherent maturity differences between cardiology, pathology, radiology, and other areas."* Stakeholders were unsure how a maturity model would be able to capture these differences accurately in a single score, suggesting the need to segment by department or use case. From a vendor standpoint, segmentation of AI maturity could also make sense: *"If you look at how the industry is separated [and selling into health systems], it's by specific disease areas [or by workflow], so separating maturity within a health system may help with engagement."*
- **The future scalability of the approach is unclear.** In the near future, healthcare IT infrastructure may change drastically—particularly for care settings outside of AMCs—leading some to wonder about the scalability of the maturity model. *"As we think about a digital future, there's an increasing openness to cloud-first infrastructure in clinics, which is different to the infrastructure in health systems. So while the AI maturity model may work in certain systems, it may not apply to others,"* one participant said. Additionally, a few participants questioned whether an AMC has the capability to implement the model broadly: *"If the goal is to assess maturity at scale, wouldn't an entity like HIMSS be better suited for implementation, given their existing reach?"*
- **Evaluating health systems alone is not sufficient.** While the proposal specifically focuses on the maturity of a health system, stakeholders also highlighted the need for a way to evaluate vendor maturity, particularly in care settings outside of AMCs. *"Eighty percent of care occurs out in the community, where AI expertise is lacking, and many assume the technologies from big companies have been vetted, but there's no real transparency from vendors. To improve the quality of AI, we need to look at the whole ecosystem, not just health systems,"* one participant said. Industry representatives also agreed on the need for both vendor and health-system assessments, noting that clear standards across these stakeholders would enable a streamlined development process that would benefit the broader healthcare ecosystem.

- **The AI maturity model should consider goals beyond benchmarking for greater impact.** Several stakeholders raised questions about the long-term goals of the AI maturity model, particularly with regard to community centers: *"Will the maturity model just reflect the state of the world? That may have some value, but if the model doesn't come with a mechanism to help under-resourced health systems, it won't address equity or access issues."* One stakeholder also shared concerns about the utility of the model for community systems during a time of financial uncertainty: *"The maturity model may be useful for community centers, but it doesn't address what they're worried about right now—addressing workforce issues and meeting clinical quality measures to obtain funding."* In short, while a maturity model may be a useful tool for evaluation, there may need to consider how the model can support the improvement of health systems, particularly those with fewer resources. Stakeholders shared that an additional goal could be aligning the maturity model with current or future regulatory guidelines from federal agencies such as the Office for National Coordinator for Health Information and The Joint Commission to tie AI maturity with standards that health systems already work towards.

Proposal 2: Best practices platform for AI governance and decision-making

The market proliferation of AI technologies has, in the eyes of some, led to flurry of activity around development and purchasing, with healthcare leaders lacking *"a trusted external entity to turn to"* for AI-related decision-making. Some stakeholders see the benefits of an ongoing AI learning network and best practices platform, as outlined below.

Overview of Proposal 2

The rapid pace of development and deployment of AI technologies has garnered countless real-world learnings for those making purchasing and implementation decisions. While these valuable lessons are often shared informally in peer networks, a community of health systems has established a structured learning network of academic and community centers that would discuss case studies and lessons related to the procurement and utilization of healthcare AI. These findings, as well as additional input from subject matter experts, operational leaders, and staff, are currently being curated into a best-practices platform that will be continually updated with evidence and case studies related to AI governance, implementation, performance monitoring, and outcomes. **The goal of the project is to create a reliable, trusted resource to guide organizational leaders on issues related to AI technologies and improve AI quality by informing better decision-making.** The additional funding from the Moore Foundation would help expand the community of health systems and its learning platform, and keep the information relevant.

The proposal aims to address the need for a trusted and well-researched resource for decisions surrounding AI technologies and can be considered analogous to UpToDate, a clinical software resource that guides providers on evidence-based practices at the point of care. However, while UpToDate is owned by Wolters Kluwer Health and requires a subscription for access, the best-practices platform would be free to the public.

Several participants noted that *"building an implementation community where people can openly ask questions and learn why something failed would be useful, especially for smaller institutions with fewer resources."* Developers also recognized the utility of such a resource for health systems, noting that a best-practices platform *"could help us all move past the hype cycle for AI"* and encourage the move toward evidence-based reasoning.

Even as stakeholders acknowledged the potential benefits of a best-practices platform for healthcare decision-makers and lauded the intent to provide it as a free resource, many expressed hesitancies around the project's potential for scalable impact for the following reasons:

- **Real-world effects of similar platforms have been difficult to measure**. While UpToDate is widely used and serves as the leading point-of-care resource for many clinicians, some lamented that *"it is really hard to tell how much of an impact it makes and whether or not it is improving care and outcomes."* Given this dynamic, stakeholders foresee that the AI best practices platform will experience similar challenges.
- **The applicability of case-study learnings may be limited.** UpToDate curates best practices based on meta-analyses from an extensive base of medical literature and opinions from leading experts. In contrast, the best-practices platform is based on findings from a small number of health systems, as well as the small but growing literature for AI, leading some to raise concerns about applicability: *"The biggest challenge is disseminating relevant information for widespread uptake across health systems. Even the best case studies will not inform you whether a particular AI technology is going to work with a certain patient population."* Additionally, while the learning network is composed of both AMCs and community health systems, a healthcare leader shared reservations, based on previous experiences, about the relevance of learnings to specific community settings: *"When we get into the details of*

> *"This proposal may establish a better model for cross-system learning and best-practice sharing than what is in existence. The question is whether that is sufficient to improve quality, especially if you can't assess the platform's effectiveness."*
>
> —Health system leader

Epic best practices, we find it challenging to implement them. Case studies are great, but they can be very site dependent; not every community system functions in the same way."

- **The rate of technological change may outpace the learning network and best practices platform.** Many participants agreed that, as one said, *"it will be hard for the group to keep up with the material"* within the cross-learning network, let alone the continuous updating of the best-practices platform. One emphasized that keeping pace may be even more challenging given the small size of the network: *"For radiology, KLAS improved quality by giving ratings on vendors and use cases. They had a dedicated team looking at data across hundreds and thousands of institutions, and that's not what we have here."*

Proposal 3: Quality-assessment standards within a specialty

A professional society is working to enable consistency of information across a specialty area to align developers, clinicians, and payers, as outlined below. While the FDA sets standards for the approval or clearance of AI technologies under the regulatory framework of AI/ML-based software as a medical device, stakeholders noted that the FDA process is *"focused on baseline safety and efficacy,"* and not on quality issues such as training data and real-world performance.[11] As a result, clinicians find it challenging to discern differences between algorithms, especially given the number of medically cleared AI products on the market. One participant offered an example: *"For detecting lung nodules, there are currently 25 different solutions. There's no way to tell which is better because it's hard to even get basic information on algorithms–the demographics on which it was trained, for example."*

Overview of Proposal 3

Developing common definitions, setting quality standards, and assessing algorithms for AI and ML-DDS technologies requires deep knowledge and expertise. A professional society, which already defines standards and evaluates performance metrics in other areas of healthcare, is proposing to lead an effort to ensure quality AI technologies by

- defining the process and performance metrics, as well as outcome measures, for standards of quality;
- curating reference datasets and collaborating with independent third parties to evaluate algorithms; and
- monitoring real-world performance of AI-technologies in effectiveness and durability, potentially with public or private data registries.

The goal of the project is to help make language and measures for AI consistent within a specific clinical specialty to set a foundation for quality improvement

across stakeholders relevant to the specialty's care delivery and quality oversight.

Health systems, self-insured employers, payers, and industry may all have internal processes and measures to evaluate AI performance, but one stakeholder highlighted the potential benefit of a professional society defining standards across the field: *"Standards compliance is not being done by the vast majority today because there's no push to do so. [The professional society] could say, 'Here are the open standards to work with us,' which could get things started."* The society's continuing work and experience in other QA initiatives also caused some to feel optimistic about the initiative's potential for success.

However, despite agreement on the need for common definitions and standards for the development and assessment of AI, stakeholders called attention to several issues to consider regarding professional societies:

- **Neutrality and objectivity.** Given that professional societies represent the interests of their membership, several stakeholders noted there may be an inherent conflict of interest involved with such organizations setting standards. *"Defining quality standards would have a huge impact across the field. I don't see how entity that represents and gets funding from clinical members can establish standards objectively,"* one said. Some were particularly concerned about transparency, suggesting that if studies show poor performance of clinicians against set standards for AI and ML-DDS, *"the professional society won't allow that study to make the light of day."* As a result, stakeholders recommended other standard-setting organizations, which were perceived as more neutral, to lead the work on quality—for example, the National Electrical Manufacturers Association, the National Institute for Standards and Technology, and the Responsible Artificial Intelligence Institute. Stakeholders also highlighted DICOM as a successful collaborative endeavor in the field of radiology.
- **Diversity.** As detailed above, professional societies can fail to represent the perspectives of non-academic clinicians. With more patients receiving care outside of AMCs, some stakeholders raised the need to consider community and retail sites as standards are established.
- **Openness to collaboration.** Certain industry representatives said that in their experience, professional societies have *"not been very friendly with efforts to co-shape processes and other standards."* Existing consortiums such as Integrating the Healthcare Enterprise (IHE) are already working on quality and performance metrics, and medical societies should consider collaborations with these groups. Indeed, the majority of stakeholders were quick to emphasize the need for collaboration, particularly for an initiative that would affect all players in an ecosystem. *"If the goal is to improve the quality of AI technologies across the field, then that requires broad stakeholder engagement. What wouldn't be good is a*

medical society setting standards, becoming the gatekeeper and then slowing down adoption of AI," one stakeholder said. As quality standards in one specialty can affect other specialties, many also raised the need to consider how multiple medical societies can work in conjunction with one another.

- **Speed.** Many participants opined that professional societies function at a pace that is not conducive to AI. One said, *"I am afraid that a society will take a year or two to update its standards and guidelines, and by that time the whole industry has moved forward."* Some believed the lack of speed is inherent to the structure of professional societies and suggested instead the need for more focused resourcing on quality issues. *"A typical society depends on academic clinicians who volunteer their time and work on projects and committees when they can, as a second job. For this initiative, we need a fully funded entity with a dedicated team of people running it,"* one stakeholder said.

The importance of incentives and collective learning across all proposals

Beyond considerations specific to each proposal, interviewees highlighted two cross-cutting themes that relevant leaders should bear in mind as the above projects advance.

Reimbursement for AI currently sits at a crossroads

First, cross-stakeholder contemplation of the realities, constraints, and resources that various players in the healthcare system face is essential. Today, AI tools are primarily implemented to increase productivity, reduce costs, and manage repetitive tasks for clinicians and health systems. As a result, providers who are rewarded for efficient practice or through value-based payment models are implementing these new technologies more rapidly. However, while such incentives drive the general adoption of AI, stakeholders agreed that, as one said, *"there's no business model"* to increase the adoption of high-quality AI. *"There's no motivation for the people creating products to validate them. It's only worthwhile if they're going to get paid, but where is that payment going to come from? That's the missing link right now. Everyone's just trying to sell with marketing pitches because there's no better way, and nobody has a strong incentive to create a better way."*

Given the lack of stimulus to propel the quality of healthcare AI, stakeholders emphasized the need for each proposal to consider the broader reimbursement landscape to address the sustainability of quality initiatives, even if views on the role of reimbursement in directly or indirectly rewarding AI in healthcare are mixed. One said, *"Thinking about payment methodology is going to be critical. Just look at the example of telehealth—it took a pandemic and a certainty about reimbursement for its adoption at scale."* Some suggested, for example, that an assessment of AI maturity could include a health system's approach to contracting, risk, and value-based payment models and that quality metrics for a specialty could align with outcomes that payers already evaluate. By proactively considering reimbursement in quality initiatives, some believe the projects could serve as case studies for the*"clear end-to-end value proposition of AI and, in turn, kick-start payment."*

That said, even payers were unclear on how payment and reimbursement for AI and ML-DDS products might evolve, making integrating a reimbursement focus into the above projects challenging. Some were open to considering innovative ways to incentivize AI to help lower costs, with one noting, *"There are plenty of financial incentives [to adopt AI] and we would love to think about new ways to look at the return on investment for these technologies."* Others were more skeptical, citing the dearth of well-designed studies and real-world evidence on outcomes in comparison to drugs. For some, there was also uncertainty about the role of payers in improving ML-DDS solutions: *"If AI enables a job to be done more efficiently and effectively with less error, that's great. But should we wrap reimbursement or value around that? I don't know."* Others were more skeptical, stating that direct reimbursement by payers for the use of AI would be an undesirable outcome. One stakeholder said, *"AI should be used as one of many tools to help improve clinical outcomes. Paying for such tools on a per-use basis, given the potential scale of these technologies, would be a bad precedent for healthcare."*

Building a foundation for quality may require an intentional structure

Even as stakeholders considered proposals on an individual basis, some noticed potential synergies between the projects discussed. As a result, many highlighted the value of multistakeholder collaboration and learning, not just within each project but across all quality initiatives as a whole. *"A lot of these forums and projects exist, but they're all operating in silos. How can we enable a centralized consortium on quality?"* one participant asked. To facilitate such centralization, some suggested the creation of a multistakeholder group—with representatives from payers, industry, and diverse health systems—to serve as a strategic advisory body across proposals *"to help ensure demonstration projects are impacting the right groups."*

Conclusion

While AI holds substantial promise and could help to solve some of the most pressing issues in healthcare today, the need for standards for development and deployment is clear. In the absence of a single entity to drive the quality of diagnostic AI in a transparent and neutral way, there are opportunities for organizations to advance the field. Specific concepts discussed include establishing an AI maturity model for health systems, formalizing a peer-learning network with the development of a best-practices platform, and establishing common definitions and standards within a specialty area. While stakeholders identified potential areas for improvement for each project, most saw value in the principles underlying all three proposals. Recommendations across the proposals aligned with the aforementioned key principles for quality initiatives in this space, specifically the following:

- **Prioritizing multistakeholder collaboration**, both within projects—through the inclusion of diverse stakeholders and partnering with consortiums with synergistic efforts—and across projects, with the potential creation of a strategic advisory group
- **Considering how quality can be affected at scale,** particularly with shifting site-of-care dynamics, the lack of resources for community centers, and the potential benefits AI may bring to underresourced populations
- **Working today with the future in mind,** given that demonstration projects are being developed in parallel with the rapid advancement of AI technologies
- **Addressing the issue of sustainability,** with consideration for how each project may impact potential reimbursement and how diverse stakeholders with competing interests can work toward a common goal of quality improvement

Many applauded the Moore Foundation's effort to push quality to the forefront of healthcare AI by funding demonstration projects. One said, *"There's a real need here, and I think Moore can really make an impact. Nobody wants to do their part because they're not getting paid; no one is putting any money in this space. Yet we have to start somewhere."*

Despite the complexities around quality for AI and ML-DDS and uncertainty around sustainably incentivizing quality initiatives, some remained optimistic about the future, particularly when reflecting on the multistakeholder engagement during conversations for this effort. One noted, *"AI quality is very important, and it will have a ripple effect on many areas of healthcare. We need more diverse collaboration, and we need to leverage subject matter experts from all areas. I think it's doable—it just needs the right structure. This forum is a great start."*

About this document

This *ViewPoints* reflects the use of a modified version of the Chatham House Rule whereby comments are not attributed to individuals, corporations, or institutions. Italicized quotations reflect comments made by participants before and during conversations relevant to this initiative.

Tapestry Networks is a privately held professional-services firm. Its mission is to advance society's ability to govern and lead across the borders of sector, geography, and constituency. To do this, Tapestry forms multistakeholder collaborations that embrace the public and private sector, as well as civil society. The participants in these initiatives are leaders drawn from key stakeholder organizations who realize the status quo is neither desirable nor sustainable and are seeking a goal that transcends their own interests and benefits everyone. Tapestry has used this approach to address critical and complex challenges in corporate governance, financial services, and healthcare.

This document is not intended to represent the particular policies or positions of the individual participants or their affiliated organizations.

Appendix 1: Participants

The following stakeholders participated in interviews and small-group discussions:

- **AGFA HealthCare:** Nathalie McCaughley, President
- **ArcBest:** Rich Krutsch, Former Vice President, People Services
- **AT&T:** Luke Prettol, Lead Benefits Strategy Consultant
- **Blue Cross of California:** John Yao, Chief Medical Officer
- **Blue Cross Blue Shield Association:** Naomi Aronson, Executive Director of Clinical Evaluation, Innovation, and Policy
- **Blue Shield of California:** Bob Plass, Medical Director, Medical Policy and Technology Assessment
- **Carelon:** Jim Perry, Vice President, Digital Care Products and Solutions
- **Covera Health:** Ron Vianu, Founder and CEO
- **CVS Health:** Kjel Johnson, Vice President, Specialty Product Strategy
- **Decipher Health Strategies:** Hope Glassberg, President
- **Duke Institute for Health Innovation:** Suresh Balu, Program Director; Mark Sendak, Population Health and Data Science Lead
- **Emory University:** Judy Gichoya, Assistant Professor
- **Evernorth:** Wiliam Lopez, National Medical Director for Virtual Care
- **GE Healthcare:** Karley Yoder, General Manager, and Chief Digital Officer, Ultrasound
- **Geisinger:** Phil Krebs, Director, Medical Policy and Clinical Guidelines
- **Google:** Matthew Thompson, Clinical Research Scientist, Health Impact
- **Healthcare Information and Management Systems Society:** Julius Bogdan, Vice President and General Manager, Digital Health Advisory, North America
- **Highmark Inc:** Matt Fickie, Senior Medical Director
- **Humana:** Jeremy Goodridge, Technology Services Principal
- **Intermountain Health:** Albert Marinez, Former Chief Analytics Officer
- **Jefferson University:** Stephen Klasko, Former CEO; Executive in Residence, General Catalyst
- **Lahey Hospital and Medical Center:** Christoph Wald, Chairman, Radiology

- **MedStar Health:** Nawar Shara, Director, Biostatistics and Biomedical Informatics; Sara Steinecker, Project Manager
- **Memorial Sloan Kettering:** Joseph Sirintrapun, Director of Pathology Informatics
- **Nuance:** Diana Nole, Executive Vice President; Calum Cunningham, Senior Vice President, Diagnostics; Sheela Agarwal, Chief Medical Information Officer, Diagnostic Imaging and AI
- **OCHIN:** Josh Lemieux, Director of Research Collaborations
- **Ochsner Health:** Phil Oravetz, Chief Population Health Officer
- **Purchaser Business Group on Health:** Emma Hoo, Director of Value-based Purchasing
- **Philips:** Sham Sokka, Former Head of Innovation and Marketing, Precision Diagnosis
- **RadNet:** Greg Sorensen, CEO, DeepHealth
- **Radiology Partners:** Kent Hutson, Director of AI Innovation, Clinical Operations
- **Roche:** Mike Bales, Director, Strategy and Operations, AI & Digital Health
- **Samaritan Health Services and Health Plan:** Sonney Sapra, Senior Vice President and Chief Information Officer
- **Stanford University:** David Larson, Senior Vice Chair for Strategy and Clinical Operations, Department of Radiology
- **Ultromics:** Helen Routh, Non-executive Director
- **University of Maryland:** Warren D’Souza, Chief Innovation Officer
- **University of Pennsylvania:** Ravi Parikh, Director, Human Algorithm Collaboration Lab
- **Vanderbilt University:** Peter Embi, Chair, Department of Biomedical Informatics and Senior Vice President for Research and Innovation; Laurie Novak, Associate Professor, Department of Biomedical Informatics
- **Weill Cornell Medicine:** Geraldine McGinty, Professor of Clinical Radiology and Population Health Sciences, Senior Associate Dean for Clinical Affairs

Appendix 2: Concepts to drive quality for ML-DDS

Stakeholders were initially asked in individual interviews to consider seven concepts to prioritize to advance the safe, effective, and fair adoption of ML-DDS:

1. Developing a framework to guide health system adoption and deployment of ML-DDS
2. Developing a framework to guide vendors/developers in developing clinical ML-DDS
3. Creating standardized procurement templates/request-for-proposal expectations for health systems to use in ML-DDS purchasing
4. Standing up a third-party accreditation service that ensures vendors meet a minimum/maximum level of performance (e.g., a seal of approval)
5. Standing up a third-party evaluator of algorithm performance and/or development process claims
6. Creating a public/private registry that monitors developer performance, potentially via use of real-world evidence
7. Creating standards around transparency expectations for ML-DDS

No single concept garnered consensus support, though specific stakeholder types tended to share similar perspectives on specific priority concepts relevant to their interests. Payers, for example, largely supported the development of data registries for real-world evidence and outcomes measurement. Industry participants saw the value of standardized frameworks to clarify how developers should operate. Clinicians and health system leaders appreciated the need for frameworks for health system development, adoption, and deployment of AI.

Appendix 3: Proposal concept review criteria

During small-group conversations, stakeholders were asked to provide input on the proposals based on the following factors:

- The potential impact of the proposals in addressing issues of standards and quality and, in turn, adoption and scalability of AI and ML-DDS
- The potential implications and unintended consequences for stakeholders outside the target audience
- The applicability of the proposals to diverse sites of care (e.g., community health systems, retail-affiliated clinics, and single-specialty centers)
- The overlaps and synergies of the proposals to existing efforts
- The interest for organizational involvement and the incentives necessary for participation

Endnotes

[1] US Government Accountability Office & National Academy of Medicine, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics* (Government Accountability Office, September 2022).

[2] Rock Health, "AI Vibe Check," *Rock Weekly*, April 10, 2023.

[3] "How FDA Regulates Artificial Intelligence in Medical Products," Pew Charitable Trusts, August 5, 2021.

[4] Andrew Wong et al., "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA Internal Medicine* 181, no. 8 (2021).

[5] Leo Celi et al., "Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review," *PLOS Digital Health* 1, no.3 (2022).

[6] Lizzy Lawrence, "The FDA Plans to Regulate Far More AI Tools as Devices. The Industry Won't Go Down Without a Fight," *STAT+*, February 23, 2023.

[7] "History," DICOM, accessed June 26, 2023.

[8] Howard Fleishon, et al., "White paper: corporatization in radiology." *Journal of the American College of Radiology* 16, no. 10 (2019).

[9] "Maturity Level Characterization of Artificial Intelligence Capabilities for Self-Assessment." AFDO/RAPS Healthcare Products Collaborative, accessed July 7, 2023.

[10] "Adoption Model for Analytics Maturity," HIMSS, accessed June 26, 2023.

[11] US Food & Drug Administration, *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)* (US Food & Drug Administration, 2019).