AHRQ National Webinar on Artificial Intelligence Tools to Improve Provider Effectiveness and Patient Outcomes – Questions and Answers
March 18, 2025

Jeremiah Hinson, M.D., Ph.D.
Johns Hopkins University
Transforming Kidney Care in the Emergency Department Using Artificial Intelligence-Driven Clinical Decision Support

QUESTION: What is the most challenging aspect of implementing an AI-informed decision-making tool?

ANSWER: The first most challenging aspect is technical - moving from a published retrospective model to implementation. For example, during the pandemic there was a period of time when there were hundreds of machine learning (ML) algorithms that were developed, but very few were implemented. That really requires having a team that collaborates with people like Dr. Hsu, who are informaticists, and really understand how you get things into the front end of an electronic health record (EHR). If you can get over that huge barrier and get things into the EHR or into an existing workflow, then the challenge becomes engaging the user in a way that allows them to understand enough without getting flooded with information. We are currently still in the phase of overcoming the concept that ML or artificial intelligence (AI) is a black box. We use a lot of technologies that are black boxes, like ultrasound images, MRI images, and CT images. I have no idea how those work, but I trust what I am seeing because there are algorithms underlying those. So, we must get to a point where people understand enough about how the tools work so that they can trust them, but not so much that they are overwhelmed with the complexity that is behind the scenes.

QUESTION: How do you think AI-informed decision support tool interventions will be spread?

ANSWER: That is a somewhat challenging question. We produce many things in academia with the intent for them to spread, but it takes significant resources for a tool like this to scale nationally. So, it really depends on getting a tool to a point where it can be commercialized and adopted. Given the complexity, circumstances may change, and it might be easier to adopt these tools locally instead. At the moment, there may be capacity for rebuilding similar tools at a local level, but for national dissemination, there is a strong opportunity for academic and industry partnerships. For example, our group received an SBIR award—Small Business Innovation Research—which enabled us to disseminate our tool through a commercial pathway. That has been one of the most effective routes for broader adoption. There are also opportunities to collaborate with other data scientists and institutions to expand into areas that have the capacity to support such tools.

QUESTION: What do you think are some areas of uncertainty in the AI informed decision support tool space?

ANSWER: That is a great question. I think one of the biggest uncertainties—building on one of the earlier questions—is how we move from having local models that perform well to models that can perform well everywhere. And I mean that in a few different ways. One relates to integration: how do you get these models implemented elsewhere, and how do you evaluate their performance when someone else is deploying them? The other is simply about model performance itself. We have seen—especially with models we have developed for acute kidney injury (AKI) or for predicting severe outcomes in influenza—that a model tends to perform best in the population where it was trained. There are several reasons for that: some are related to the patient population, others to the practice patterns of the hospital. These models often need to be retrained locally to achieve optimal performance. Figuring out how to do that within an appropriate regulatory framework is a little bit of a challenge. Right now, the regulatory framework that is going to be applied to them does not really allow for that kind of local adaptation, which can result in suboptimal algorithms. So, this is a very uncertain area, and everyone is learning as we go. I expect a lot of development in this space over the next few years.

QUESTION: How do you best get buy-in when wanting to place an AI tool in a provider's workflow and who do you need buy-in from in order to place AI in the clinical environment?

ANSWER: The brief answer is that you get buy-in before you develop the solution. The best way to get buy-in is to involve the end-users during the development of the tool, and then they become a champion of the tool because they understand what it is. If that is not possible, then you find an influential individual within the group and then you train them with the tool. You really teach them how the tool works and then let them share that with their peers. Those have been the two most effective methods for us.

QUESTION: I have attended other webinars that warned of risk for bias in algorithms and there has been concern about that in other kidney disease studies. How can we (the end-users) trust that the algorithms used to support the AI tools are addressing bias?

ANSWER: We train our algorithms in a broad population and perform validation across the entire cohort, and also in sub-cohorts, divided by race and ethnicity to ensure there is no discrepancy in performance between groups.

QUESTION: What is the process for reviewing the algorithms to ensure they are up to date based on evidence-based practice, and how often are they reviewed?

ANSWER: The algorithms predict onset of acute kidney injury based on consensus definitions and are trained using local electronic health record data. Unless the definitions for kidney injury change, there would really be no impact of evolving evidence on the algorithms' operation. In terms of monitoring performance, it is essential to monitor the function and predicted performance of algorithms overtime at each site where they are deployed and is something that we do regularly.

QUESTION: Any thoughts on the recent paper "Low Responsiveness of Machine Learning Models to Critical or Deteriorating Health Conditions" by Tanmoy Sarkar Pias

et al.? I came across the headline: "AI health models may miss critical conditions — AI systems designed to predict patient mortality are failing, with models missing about 66% of critical health conditions leading to patient death," according to a study published in Communications Medicine. Study co-author Danfeng Yao emphasizes that medical knowledge needs to be integrated with patient-care technology, and that training AI models solely on data is insufficient.

ANSWER: We certainly agree that ML models are not sufficient to operate in isolation, but instead, should be used as decision support for human clinicians. There are now, and will continue to be for some time, a multitude of patient-level factors and events that are not captured in an electronic format and are unavailable to ML models - making the insights of human operators critical. Also, most ML output is probabilistic and as with other diagnostic modalities, will never be perfect in its performance. AI/ML is just another tool to help us detect trends earlier and make better decisions on the whole.

Jessica Keim-Malpass Ph.D., A.P.R.N.
University of Virginia

Predictive Monitoring: IMPact of Real-Time Predictive Monitoring in Acute Care Cardiology Trial (PM-IMPACCT)

QUESTION: What are some lessons learned from implementing the CoMET system and integrating it into the workflow?

ANSWER: So, I have alluded to some of the lessons learned, particularly around methodological considerations, but I really cannot emphasize enough the importance of working at a partner level with end users from the very beginning. This includes involving a wide range of clinician types and end users early on, especially nurses or other clinicians who are often the first touch point for the system. We worked extensively with them because there were trust issues. Any kind of analytic tool like CoMET is going to have a relatively low positive predictive value. For example, if we had set an alert threshold at a value of two—which we did not—the positive predictive value would have only been around 30%. That means most of the time, a signal like that would not correspond to an actual clinical deterioration event, particularly given how rare such events are in this patient population. So, it is really about using terms, thresholds, and tradeoffs in ways that help clinicians understand that not every alert or rise indicates something bad is going to happen. Rather, it should be integrated with other ways of thinking that clinicians already use—we all have our own prior probabilities. These tools should be used to augment what we already know.

Where I think it can be especially helpful at clinical junctures—when you are unsure and need some objective data to support your gut feeling. That is where these tools can really add value. Trust and collaboration are key to understanding how clinicians want to use these systems. In fact, our decision not to implement an automatic alert was directly based on feedback from nurses. Had we not engaged them, we might have implemented alerts that could have led to alert fatigue and increased burnout.

QUESTION: While the CoMET system models were trained on clinical deterioration events, have they been used in other off target ways? And what are clinician's perspectives of this?

ANSWER: This is one of my favorite aspects actually and one of the ways that we want to extend this work in the future. Very early on, bedside nurses told us how they were using CoMET to make decisions that had nothing to do with clinical deterioration. One of our nurse practitioners said, “I look at CoMET and I can figure out if I have time to go get coffee before I start or not.” We know that pain and sedation, particularly in the ICU, and other elements have these autonomic signatures that CoMET is able to pick up on, because of some of the EKG computations. So, the next kind of push we are doing, and this is completely stakeholder driven, is figuring out ways that nurses, physicians, respiratory therapists, and pharmacists like to use CoMET that are not based on clinical deterioration.

This can be problematic in some ways because if the model is not initially trained on whatever event you are assessing, you have the risk of it not performing well or not really getting at what you are intending. I am looking forward to using it to better understand supportive care needs for some of our patients. This is in line with thinking about how to refine the cohort going forward. I am really excited to think about this in the context of heart failure patients in particular, and understanding responsiveness to therapy, which is not something that we have done yet but is hugely important. So, what if we treat these physiological scores like a pharmacodynamic marker? What is to say that the tool cannot be used to assess in real-time how a patient is responding to a titration or an escalation in care. And so, those are the things that I am most excited about and partially because of my background but also because this is what our bedside nurses are telling us too.

QUESTION: What do you think are some areas of uncertainty in the AI-informed decision support tool space?

ANSWER: I think we may need more nuanced methods to understand what we should do with it. For diagnostics, the path is more straightforward. For ongoing time series models, it was more challenging—but I do believe reinforcement learning, discrete event simulation, and other methods could help guide us. These approaches can guide clinicians in understanding when they should act versus not.

QUESTION: How do you best get buy-in when wanting to place an AI tool in a provider's workflow and who do you need buy-in from in order to place AI in the clinical environment?

ANSWER: It is important to highlight a diversity of clinician types, targeting everyone that the analytic will potentially impact. Also, integrating key infrastructure folks on the data side of the health system, which is complex and there are so many different people involved with implementing any kind of analytics. The last thing I want to say is that one area that has not been touched is patient and family integration in terms of model development and implementation and this is the next frontier. I will speak for myself, but we did not include patients and families. Prior to my involvement in this modeling group, there was no nurse representation. So, thinking holistically about end users, including

patients and families, both in terms of the things that they care about and the ways that the information impacts their entire care pathway.

QUESTION: What is the reference for development of the CoMET scoring system and description of the scoring method?

ANSWER: Details can be found in the following publications:
Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, Moorman JR. Signatures of Subacute Potentially Catastrophic Illness in the ICU: Model Development and Validation. Crit Care Med. 2016 Sep;44(9):1639-48. doi: 10.1097/CCM.0000000000001738. PMID: 27452809; PMCID: PMC4987175.

Keim-Malpass J, Moorman LP, Moorman JR, Hamil S, Yousefvand G, Monfredi OJ, Ratcliffe SJ, Krahn KN, Jones MK, Clark MT, Bourque JM. Prospective validation of clinical deterioration predictive models prior to intensive care unit transfer among patients admitted to acute care cardiology wards. Physiol Meas. 2024 Jun 5;45(6). doi: 10.1088/1361-6579/ad4e90. PMID: 38772399.

Keim-Malpass J, Moorman LP. Nursing and precision predictive analytics monitoring in the acute and intensive care setting: An emerging role for responding to COVID-19 and beyond. Int J Nurs Stud Adv. 2021 Nov;3:100019. doi: 10.1016/j.ijnsa.2021.100019. Epub 2021 Jan 5. PMID: 33426534; PMCID: PMC7781904.

QUESTION: Very interesting findings. I would be interested to understand what potential additional insights or differential diagnostics might help avoid unnecessary transfers if such diagnostics were available be it lab results or point-of-care ultrasound (POCUS) findings (such as echos acquired by bedside nurses interpreted by intensivists or cardiologists remotely) that could be used by an emergency response team (ERT).

ANSWER: I certainly think that the intersection of advanced analytics coupled with diagnostic techniques is the next frontier. Targeting who may need more advanced diagnostic recognition is certainly a use case for this analytic.

QUESTION: Outcomes are impacted by continuing care - have you planned or implemented changes to support continued care? Such as opioid use disorder, or breast cancer biopsies, or to retain patients in your network?

ANSWER: We are currently assessing heart failure outcomes and outcomes synergistic with continuity of care for our next analysis.

QUESTION: Given the low image and population data of minority groups, how do you address data gaps?

ANSWER: Bias recognition in model development and deployment are also the next frontier of AI in medicine. We do not use demographics in our models, but to your point, this does not mean that bias does not exist in our models.

William Hsu, Ph.D.

University of California, Los Angeles

Using Artificial Intelligence to Support Efficient Same-Day Diagnostic Imaging in Breast Cancer Screening

QUESTION: Information from simulation studies is only as good as how accurate the simulation is in modeling the real world. How would you validate the RSS scheduling algorithm in AI-driven, same-day diagnostic triage in a clinical environment?

ANSWER: That is a great question. Whenever we do these types of models, your model is as good as your assumptions. There is always that quote that "all models are wrong, but some are useful." I think in this case, some of what we tried to do was test our assumptions. This included working closely with our clinical partners to ensure we were accounting for potential factors that could influence the model. We also conducted studies to collect data on how the workflow actually occurs, to help parameterize the model, essentially putting the right numbers in to make it accurate. Ultimately, we ran retrospective analyses, comparing the model's output to real-world data to assess how accurate or aligned they were.

That said, models inevitably rely on many assumptions, so there's definitely room for improvement. For example, questions have been raised about how risk-stratified screening paradigms could affect underrepresented populations, and whether the model's assumptions hold over time. These are ongoing areas of focus for us. My hope is that through regular validation and close collaboration with clinicians, we can build a model that supports better evidence-based decisions in practice.

QUESTION: What are strategies for using information from an external validation study to improve the implementation of an AI algorithm at a target institution?

ANSWER: When we do these external validation experiments, it certainly takes a good amount of effort to curate the data and run the actual analysis. But it gives us valuable insights into how much we can trust the AI algorithm. Of course, in the commercial space, when you look at different AI algorithms, everyone claims theirs is the best and highlights its benefits, but those claims can vary, especially when you try to run the algorithm on your own population. This is particularly true if your population differs significantly from the one the model was originally trained on, which you may not even know, since vendors are not always fully transparent about their training data.

That is why external validation is so important—it helps you understand how well the algorithm actually works in your specific clinical setting. Metrics like sensitivity and positive predictive value can tell you how many cases the AI might flag, and how many of those you would actually need to follow up on. These are important considerations when discussing implementation with clinicians or others who will be using the AI directly. We need to clearly understand the trade-offs—such as dealing with more false positives in exchange for potentially catching cancers earlier. So, having stronger evidence from external validation can really help guide those conversations before rushing into implementation.

QUESTION: What do you think are some areas of uncertainty in the AI-informed decision support tool space?

ANSWER: The current regulatory framework prevents us from doing local adaptation of these algorithms, which could potentially optimize their performance. But another important aspect is figuring out how to optimize the human-computer interaction. Oftentimes, when we talk about the current phase of AI, there have certainly been many technical advances, but this is not entirely new. In breast screening, for example, AI has been around before—there was a previous generation of computer-aided diagnosis in the 1990s that ultimately failed. That failure was not just about the algorithm itself, but also about how humans interacted with it, which in some cases actually reduced provider performance. As we move forward with implementation, we should learn from those historical mistakes. Ultimately, I think many of the points we discussed today are helpful tools for avoiding those kinds of pitfalls.

QUESTION: How do you best get buy-in when wanting to place an AI tool in a provider's workflow and who do you need buy-in from in order to place AI in the clinical environment?

ANSWER: In the past, what I have seen is that those who were the most enthusiastic were the ones who championed purchasing and implementing AI and that is how it got adopted. But over time, we have realized there are many downstream impacts, and we cannot have everyone simply purchasing their favorite AI algorithm. There needs to be governance and rigor around how we choose AI tools and evaluate how they function within our health system.
More and more, at least in our health system, there is a formal process in place for evaluating tools, their use cases, why they are being used, who will be impacted, and what the broader implications are. I think that is really the right approach: to understand and anticipate these issues. But ultimately, the buy-in still comes from the enthusiasm and the recognition that AI has the potential to truly transform the tasks we perform.

QUESTION: Are you familiar with AI for quick, non-clinical translations?

ANSWER: This is a very active area of research, and while many tools exist (mostly commercial), their accuracy has not been rigorously evaluated. People have used ChatGPT as a generative tool for language translation (see https://jamanetwork.com/journals/jama/fullarticle/2823653 for an example). There are other commercial tools if you search for medical language translation. For a more detailed technical resource, I would recommend looking at https://github.com/AI-in-Health/MedLLMsPracticalGuide. There is a dedicated section on Medical Language Translation.

QUESTION: Is the AI reimbursable by insurance? Is a note put in the patient's chart to signify that AI is utilized?

ANSWER: There are some classes of AI algorithms that are reimbursable by insurance. CMS Coverage: The Centers for Medicare & Medicaid Services (CMS) has started to provide coverage for certain AI-specific Common Procedural Terminology (CPT) codes. CMS has also created New Technology Add-On Payment (NTAP) designations for AI devices, which provide temporary reimbursement for new technologies. This question

warrants further discussion, but I would point you to the following paper: https://www.sciencedirect.com/science/article/pii/S1546144023005963.

QUESTION: What is the process for reviewing the algorithms to ensure they are up to date based on evidence-based practice and how often are they reviewed?

ANSWER: I am unaware of a broad effort to review algorithms for being up-to-date based on evidence-based practice. Much of this is currently left to local governance (e.g., committees at institutions that oversee the implementation of an AI algorithm locally). As an example, societies such as the American College of Radiology recognize this as a gap and have started to propose certifications for safe and effective AI use e.g., https://www.acr.org/Data-Science-and-Informatics/AI-in-Your-Practice/ARCH-AI.

QUESTION: Are these types of AI tools generally developed within a health system or spearheaded by EHR vendors?

ANSWER: Algorithms can be developed by EHR vendors, commercial entities, and in-house teams. However, EHR vendors and commercial entities already integrated into the institution's health information technology environment have the easiest pathway toward deployment.

QUESTION: Outcomes are impacted by continuing care - have any of you planned or implemented changes to support continued care? Such as opioid use disorder, or breast cancer biopsies, or to retain patients in your network?

ANSWER: Our same-day diagnostic service is designed to support the patient throughout the diagnostic and treatment process. As part of our simulation modeling, we examine how providing same-day diagnosis of these patients will impact downstream care.

QUESTION: Any thoughts on the recent paper "Low Responsiveness of Machine Learning Models to Critical or Deteriorating Health Conditions" by Tanmoy Sarkar Pias et al.? I came across the headline: "AI health models may miss critical conditions — AI systems designed to predict patient mortality are failing, with models missing about 66% of critical health conditions leading to patient death," according to a study published in Communications Medicine. Study co-author Danfeng Yao emphasizes that medical knowledge needs to be integrated with patient-care technology, and that training AI models solely on data is insufficient.

ANSWER: Thank you for bringing this interesting article to my attention. I agree with the authors' emphasis on evaluating and understanding the AI's performance and reasoning in critical applications. We must ensure the external validity of an algorithm (that it performs correctly on the target population). In my own work, I have shown that breast AI algorithms also have mixed performance, underscoring the need for continuous validation (see https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2798829 and https://academic.oup.com/jbi/article/7/1/16/7821505). Moreover, algorithms may also degrade in performance over time. We take into account that AI algorithms are imperfect in our simulation studies, which helps us plan how to ensure that the

radiologist is always confirming whether the individual requires a diagnostic exam (or reviewing negative exams at a later time to ensure that a cancer is not missed).

QUESTION: Given the low image and population data of minority groups, how do you address data gaps?

ANSWER: One of our goals is to make breast screening more accessible to patients by providing diagnostic care in a single setting rather than requiring a callback and additional appointment. We feel this paradigm, if it can be scaled, would be helpful in rural communities where getting to a center that can perform breast biopsies is challenging. Accommodating individuals on the same day can address access issues, such as eliminating multiple trips to/from the imaging center.

QUESTION: What are the economic benefits of your study? Did you do a cost benefit analysis? Is it worth the clinician's time to use this tool?

ANSWER: We are actively conducting a cost-benefit analysis as part of our simulation studies. While changes in the breast screening workflow may introduce new short-term costs, such as the need for additional diagnostic imaging capacity or an additional radiologist to handle screening same-day diagnostic exams, we also believe that AI can help mitigate these costs. We also anticipate that by performing diagnostic exams immediately rather than requiring them to be scheduled at a later date, we will ultimately be able to increase the capacity of the imaging center over time.