February 2024 Info Session/AI in Quality Measurement

Q & A Summary

What about the studies that show AI is giving false results and what has been done to review biased algorithm?

One concept is to utilize Assurance Labs to ensure models are safe and effective. Assurance labs are a critical part of testing models and identifying if the model is giving an incorrect answer. Assurance labs can also inform limitations for those models as well as test and offer independent validation.

The other core part is to test the model locally for validation to drive quality improvement. In other words, how is the model performing on our patients?

Where do you see generative AI helping with quality measurement in the future?

It's an emerging area and a lot is unknown. Clinical quality measurement is a great use case for these models. We have a lot of infrastructure already developed for human review (e.g., technical expert panels). There is a lot of potential in three areas of clinical quality measurement: 1) Expanding our knowledge about harms through large language models, 2) Identifying mechanisms for improved performance, and 3) Expanding our knowledge of context since these mechanisms are performed under different settings.

In the long term, we may see using large language models to populate standards. There may be opportunities to make connections between disparate data elements and creating standards for data extraction.

How do we prevent discrimination (like insurance coverage) when this data is available?

There is definitely some tension between the basis of insurance (which relies upon the concept of pooled risk) and the more precise predictions of risk that AI enables. This tension is not unique to health care but applies to property and casualty or other forms of insurance. But the experience of many AI researchers is that AI algorithms often seem to be less biased because of the consistency in application. But in the end, we need to not consider AI algorithms in isolation but AI as a component of a human-AI system, and to develop approaches to assure such systems to not have discriminatory impact.

What are the data quality risks associated with these methods, especially for unsupervised methods?

Again, in many respects AI is more robust to data quality problems than traditional methods, including missing data or data entry errors. AI may be able to more readily identify and account for such anomalies or even systemic patterns of such risks.

Could the notion of showing providers a profile for "patients like this" create a bias against diagnostic procedures/lab work (i.e. if we don't test for an issue, that data isn't part of the patient's record)?

This is known as the dark field effect – If you aren't testing or asking for it to be done, you won't have the data to inform those models. This is a challenge as we move forward in the use of AI. We are essentially going to be creating data sets that reinforce certain pathways (model fit). The data being generated is taking individuals down some paths and not others. There is no easy solution, but we need to periodically look at different datasets that haven't been exposed to train the model. This will help address the potential problem of "model overfit."

How do LLMs reduce the burden for measuring healthcare quality?

Large language models (LLMs) definitely have the potential of reducing the burden of data collection. For example, LLM may be used to summarize and extract relevant data from narrative forms, allowing patients and clinicians to engage verbally and freeing the clinician from entering data into forms on a computer.

When the data is flawed coming in how can the learning be correct? For example if the electronic blood pressure measurements are incorrect or with verbal tools not obtaining the correct words. How can AI give accurate predictions?

As mentioned above, this is potentially one of the ironies of AI that such algorithms may be more robust to flawed data. The proposed "model cards" that describe the data upon which the model was trained and the results of that training are definitely part of the solution, allowing for transparency and traceability.

What is the interaction or intersection between healthcare data, data standards and coded terminology and LLMs? Would mapping to coded terminology still be necessary, or would this be performed by the LLM?

In the future, LLMs and the AI agents developed from them will likely handle mapping of unstructured data to coded terminologies.

It seems like an AI model could work as a natural language processing program to extract clinical information from narrative text. This has the potential to decrease clinician burden in quality measurement. Thoughts?

Absolutely. If AI can't help us with tasks like that then what good is it!

Seems like the USCDI is a key component of allowing eCQMs to be standardized and "calculated" and for AI models to be deployed at scale in both individual patient hands and within hospital systems. What efforts are being undertaken to compel the advancement of this standard. It doesn't seem like we'll see the requirement for v3 until early 2026.

CMS has committed to full digital measurement, including the use of USCDI, by 2030. See the CMS National Quality Strategy for more information.

How do we validate AI output in quality measurement?

The same way we validate measures in general. AI output is "input" into a quality measure, and we are concerned with the reliability and validity of such inputs. Because AI is automated, we are less concerned with reliability; however, we need to ensure that the same AI system applied in different contexts but to the same underlying facts yields the same output. We will still need to compare AI output to a "gold standard." The challenge will be when AI itself is used to generate the gold standard. In that case we will need to rely upon the attestation to an assurance process.

Regarding inequities, are there plans to leverage gaps in data to geographically direct engagement efforts specifically toward areas where a paucity of data is observed?

One of the other additional ironic benefits of AI is that AI is very democratic, and will enable data collection in circumstances that were previously too burdensome.

How do you test the AI "model card" for CQMs?

Apply, apply, apply. When experience indicates that the CQM is not having the impact we anticipate, we need to be able to review the model card structure to determine whether any additional information might have prevented that result.

Are all these models created by CMS or are they handled by companies in the AI cutting edge (like Google, etc.)? Where can we find more information on the model cards? Also, are there any existing models to leverage on Github?

There are hundreds of model developers, which is why model cards are so important. A “registry” of model cards is likely necessary. To date, no organization has stepped forward to provide such a registry and Github like repository.

Do you see AI being used with the nationally integrated data in TEFCA?

AI will be used in every repeatable process, including data exchange.

A slide was shown depicting the pathway from CMS to move to digital quality measures using FHIR. However, the slide was labeled as outdated due to advancements in quality measurement with AI. Does this mean we are not moving towards digital quality measures because there is a new path?

The path to digital measures is the same, and the standards are the same, but AI will be increasingly used in the implementation of that pathway and those standards. See the CMS National Quality Strategy and the eCQI Resource Center on dCQMs for more information.

Where can I find the slides from today’s session?

Slides for this MMS Info Session and are posted to the MMS Hub here. Slides for all MMS Info Sessions may be located under Educational Resources.

Where can I learn more about the Coalition for Health AI?

More information can be found at https://coalitionforhealthai.org/.