Astrata

Read More....

Search

Search

Tags

Accuracy (1) Careers (2) Data (7) NLP (9) Prospective HEDIS (1) Providers (1) Surveillance Measures (1)

Real World NLP Validation for Quality Measurement

Feb 14, 2022

NLP is gaining traction in both payer and provider markets because it can markedly reduce the extraordinary labor costs associated with measuring healthcare quality. I'm often asked about how we validate ou NLP systems for measuring quality for programs like HEDIS® and MIPS. The Astrata Team put together this FAQ to help educated buyers understand the right questions to ask. Enjoy!

Rebecca Jacobson, MD, MS, FACMI
President, Astrata

How is the accuracy of the NLP measured?

NLP “accuracy” (often termed NLP performance) is measured using two specific performance metrics: Precision and Recall. If these names sound strange, it might be because they come from the information retrieval and computer science community. Precision is also called Positive Predictive Value (PPV), and Recall is also called Sensitivity. PPV and Sensitivity are usually more familiar to quality and measurement experts. Sometimes, you may see a third metric called the F1 (or F measure). In most cases this is a mean of Precision and Recall, under the condition that they are weighted equally.

		Gold Standard	
		True	False
System	True	TP	FP
	False	FN	TN

Precision (PPV) = TP/(TP+FP)
Recall (Sensitivity) = TP/(TP+FN)

What do Precision and Recall actually measure?

Precision is a measure of False Positives and Recall is a measure of False Negatives. For the measurement nerds you’ll want to check out the definitions in Box 1 above.

Let’s get practical and imagine an NLP system that evaluates a set of EMR charts for members that are in your gap list. The NLP system classifies these cases as either *Hits* or *True Gaps* – a binary choice. A *Hit* represents a case where there is sufficient evidence in the clinical notes to determine that the member is already compliant with a specific measure despite the absence of a claim, and a *True Gap* represents a case where the documentation suggests that the member is truly

abstractors do less unnecessary work, and the risk of reporting data incorrectly goes down.

Recall tells you the percentage of cases labeled as *True Gaps* by the NLP system, that are in fact *True Gaps.* If this number is less than 1 (or 100%) it means that there are False Negatives – cases where the NLP system calls something a *True Gap* but an abstractor would disagree and call it a *Hit*. As recall increases, there is less risk that you miss cases that could count towards your overall compliance rate.

We'll discuss more about how you know whether something **really** is a Hit or Not Hit when we discuss Gold Standards.

Do Precision and Recall help me determine product value?

Yes and no. High Precision and Recall are important pre-requisites, but they are not sufficient to quantify value.

Why can't I use Precision and Recall as a proxy for product value?

Precision and recall are very important in validating an NLP system, but they don't tell you about the value of the product. That's because the value of the product depends on what impact the product makes on key performance indicators such as efficiency. An NLP system could be perfectly accurate (Precision = 1, Recall = 1), and yet the use of the product provides no immediate value over the standard practice. Imagine a measure where 100% of your gap list is compliant, the population is small, and it is simple and fast to abstract each gap in the EMR. NLP may provide no value in triaging cases, and no value in speeding abstraction.

How is the value of an NLP product measured?

This is a much harder question because it depends on the task that the NLP is being used for. In the case of quality measurement, we can imagine several aspects of value including (1) the reduction of labor costs, (2) the ability to move medical record review to the measurement year (often called prospective or

last-pass quality review after your abstractors are done, then you are probably most interested in the potential increase in your rates by finding missed compliant cases. On the other hand, if your system is a first-pass triage system (like Astrata's Chart Review) then reducing your labor costs might be your most important value proposition. In fact, when NLP is used for prospective HEDIS, most payers are trying to keep costs down as they transition to prospective HEDIS review while the HEDIS sample is de-emphasized and eventually discontinued.

Hits Per Hour

The number of hits or exclusions closed by the abstractor, divided by total time.

Speed Up Factor

Multiple by which NLP-powered abstraction increases efficiency over standard practice.

For the task of prospective HEDIS, we usually want to calculate two things. First, we want to know how quickly the medical record abstractors can close hits and exclusions – the gap closures per unit time. We call that *Hits per Hour*. That's different than knowing how many cases abstractors can review. In prospective review, you won't expect to look at every case. Your goal is to focus abstraction on the cases that increase your reported rates to get them closer to your true rates. Only hits and exclusions are of value here. Second, we want to compare this value to what you can achieve using your standard processes. We call that ratio the *Speed Up Factor*.

makes prospective, population-based medical record review feasible and cost effective.

That may be true, but what if Precision and Recall are low? Does that say anything about product value?

While high Precision and Recall do not necessarily equate to high value, low precision and recall definitely hint at reduced value. At a Precision of 0.5 (or 50%), you are essentially tossing a coin as to whether the member is a *Hit* or *True Gap*. Your abstractors would be looking at many more negative cases. This will reduce the efficiency gain, producing a decrease in closures per hour as well as a reduced speed up factor.

How are Precision and Recall usually measured?

While there are several ways to measure NLP performance, the most rigorous way involves the use of independent, expert annotators to perform the same task that the NLP system is trying to perform. The humans make their judgments first, and then the system is then measured against this “gold standard”.

In our example system above, we would ask multiple trained medical record reviewers to review a set of medical records and determine whether it is a *Hit* or *True Gap*. These human judgements are based on annotation guidelines that derive from the HEDIS measure definitions. And we call that expert-labeled set a *Gold Standard*. Once the dataset has been labeled in this way, we run the NLP system on the same data. We then determine where the NLP agrees with the *gold standard* that a case was a Hit (True Positive) or a True Gap (True Negatives) and where the NLP system incorrectly labeled something a Hit (False Positive) or a True Gap (False Negative). These values are then used to calculate Precision and Recall as shown in Box 1. One benefit of a gold standard is that you are able to measure both Precision and Recall.

Is recall important in prospective HEDIS? There is no way I can close all the positive cases!

There is no doubt that prospective, population-based HEDIS is very different

team may have so many hits and exclusions to close, that they never actually get to the end. In this case, precision becomes the critical metric to watch.

What happens if NLP becomes a method for generating standard supplemental data? Does this impact the performance measures I should care about?

Currently, NLP can only be used in an assistive mode to drive reductions in labor, or to find missing gap closures. But NLP performance is becoming quite good and now rivals human performance. If NLP systems were to become a method for generating standard supplemental data, then the key performance metric to pay attention to will be Precision. That's because we want the quality of any data being reported to be very, very high. As we transition to automation, we can expect that some fraction of cases will continue to be reviewed by human abstractors. These might include the more difficult and nuanced cases, or cases that NLP systems cannot classify with high confidence. During that time, systems that present their low confidence cases to humans will not have to worry as much about recall. As the NLP systems improve, we can expect more cases to be automated and fewer cases to be read by humans. At this point recall becomes much more important to ensure that we do not underreport measure rates.

Do you have to use a gold standard? Can you measure NLP performance in other ways?

There are other options beside using a gold standard which can still give you a lot of useful information. For example, *expert review* of NLP output can be help you determine Precision (but not recall). When using *expert review*, you provide a set of cases that an NLP system has already classified and asks one or more experts to tell you whether it was right or not. Unlike a *gold standard*, this method does not start with a random set of cases labeled independently by the system and the experts. An advantage of *expert review* is that it can often be done at much lower cost than developing a *gold standard*. Some people argue that *expert review* is more biased than using a gold standard. But with the right type of experts and the right review workflow, it is possible to minimize that bias of these evaluations. One disadvantage of using the *expert review* method is that

Another type of evaluation is to monitor the *percent agreement* as real-world abstractors use the NLP output. This is the percentage of times that your abstractors agree with the NLP system when it serves up a hit or exclusion for them to close. Percent agreement is usually a good estimate of Precision in the wild, as long as you keep in mind that medical record abstractors can make mistakes too.

Are Precision and Recall constants for a given system, or do they change over time and across data?

Precision and Recall are definitely NOT constants, and you should be wary of anyone who tells you that their system will always produce the same performance metrics on new data. NLP systems are known to have portability issues, meaning that when they are moved to a new setting, performance degrades. There are many reasons for this (I feel another blog coming on), but the important thing to understand is that you won't necessarily know the performance on your data unless you or someone else measures it.

Does that mean that measuring the system against one gold standard won't necessarily predict the performance on my data?

Correct. The ability to generalize from performance metrics obtained on one gold standard to your environment will depend on how similar the data is in your environment when compared with what's in that gold standard. Although it's tempting to try to come up with a dataset that is so diverse that it represents all data – it is simply not practical or likely to work. Measurement of NLP performance really has to be a local phenomenon.

Will NLP performance and value metrics vary by measure?

Precision and Recall will absolutely vary by measure within a given population, and so will the value metrics like hits closed per hour and speed up factors. It's important to make sure that evaluations are being done for each measure that you plan to use.

What is the unit of analysis for NLP performance and why does it matter?

unit of analysis, we are making our judgements about what's right or wrong for each document. Which means that if there are two or three individual statements that could be used as HEDIS evidence, it doesn't matter if the system finds one of them or all of them. When we evaluate with mention as the unit of analysis, we are making our judgements about what's right or wrong for each individual mention of HEDIS evidence within a document. Mention level analysis is the most rigorous, but also the most expensive and time consuming. On the other hand, member level analysis can be too coarse in most cases. When measuring NLP Performance for Quality, Document is actually the unit of analysis that fits quality measurement tasks most closely. Astrata measures it's NLP performance metrics (precision and recall) at the Document level, but tracks statistics such as % agreement at the member level.

Want to learn more? Get in touch!

ASTRATA, INC

PRODUCTS & SERVICES

- Louisville, KY
- Pittsburgh, PA ✧
- Raleigh-Durham, NC
- Toronto, Canada

CONTACT

- Info@astrata.co
- Follow us!
- Privacy Policy

Contact Us!