Published on *The National Law Review* https://www.natlawreview.com

Unpacking Averages: How Accurate Do Class II Medical Devices Need to Be to Obtain 510(k) Clearance?

Article By:

Bradley Merrill Thompson

It's common for a client to show up at my door and explain that they have performance data on a medical device they have been testing, and for the client to ask me if the performance they found is adequate to obtain FDA clearance through the 510(k) process. I often respond, very helpfully, "it depends." But for some reason clients aren't completely satisfied by that.

I then volunteer that a general rule of thumb is 95%, but that this is just a rule of thumb. For Class II medical devices undergoing review through the 510(k) process, the legal standard is that the applicant must show that the device is "substantially equivalent" to devices already lawfully on the market. It's not a real precise standard. But recently I wondered, what do the data say regarding cleared medical devices? Answering that question is the focus of this post.

Big Caveats: Reader Beware

Normally in these posts I like to give you my results upfront, and then start explaining them. But before I do that this month, I'm going to give you some big caveats on this.

The biggest caveat is that there is no central database to answer this question, but rather informally written 510(k) summaries that can be accessed in PDF form. In data science, we call it unstructured text because, well, it's disorganized. It's just free text. And it is not even well-organized free text. Thus, it was difficult this month to do the work using natural language processing techniques to extract the relevant percentages from the text. More on that later.

I want to offer a particular forewarning to any engineers among you readers. You will likely be frustrated with this analysis because it is imprecise. Fundamentally, it is an analysis of English text. You will find yourself asking, exactly what does this analysis measure? The honest answer is it measures the frequency of certain statistics associated with certain words such as "accuracy" in 510(k) summaries. That's it. It's not any more precise than that. Thus, if the author of a 510(k) summary happened to go off on a tangent about the "accuracy" of political polling in the United States, those data would be in here. The only saving grace is that I don't think this happens too often. But the authors of such summaries could, more conceivably, for example, talk about the accuracy of the predicate device. I would just point out that's not entirely irrelevant, though, to our task of analyzing the accuracy of 510(k) cleared devices.

Not only is the analysis imprecise, it's also likely biased. The biggest source of potential bias is that FDA doesn't require everyone who writes a medical device 510(k) summary to include the results of accuracy testing that may have been required. In a sense, the data shared are only those volunteered by the manufacturer. It seems intuitive that the bias would be toward those who tend to have higher performance results because they are willing to reveal that performance publicly.

Further, performance testing isn't required for a very large number of 510(k) submitted to the agency. If a new medical device is descriptively substantially equivalent, meaning it has much the same intended use and pretty much the same design and technical features, there is no need for any performance testing at all. And thus, by extension, performance testing isn't included in many 510(k) summaries.

Indeed, a 510(k) summary itself is not always required, in that some companies instead choose to include a statement that they will make available their entire 510(k) submission in lieu of providing a summary. It's not terribly common. But if a company didn't want to write a 510(k) summary, they don't have to.

On the whole, I only found what I call performance testing data in about 570 510(k) summaries for the years 2001 through May 2023. Thus, I submit that these results must be taken with a huge grain of salt. Eyes wide open.

Results

Here, in graph form, are the results. I have grouped the resulting percentages in intervals of 5%. In other words, the percentage 95.78% is in the bucket for 95% to 100%.

Explanation

My source for the data is the 510(k) summaries available on FDA's website in the 510(k) database.

510(k) Summary

FDA's regulation at 21 C.F.R. § 807.92(b) specifies the contents of a 510(k) summary, and in particular the performance information required, as follows:

510(k) summaries for those premarket submissions in which a determination of substantial equivalence is also based on an assessment of performance data shall contain the following information:

(1) A brief discussion of the nonclinical tests submitted, referenced, or relied on in the premarket notification submission for a determination of substantial equivalence;

(2) A brief discussion of the clinical tests submitted, referenced, or relied on in the premarket notification submission for a determination of substantial equivalence. This discussion shall include, where applicable, a description of the subjects upon whom the device was tested, a discussion of the safety or effectiveness data obtained from the testing, with specific reference to adverse effects and complications, and any other information from the clinical testing relevant to a determination of substantial equivalence; and

(3) The conclusions drawn from the nonclinical and clinical tests that demonstrate that the device is as safe, as effective, and performs as well as or better than the legally marketed device identified in paragraph (a)(3) of this section.

Frankly, FDA's regulation is general and does not specifically require any particular performance metrics be stated. As a result, many companies do not voluntarily include that. Rather, they simply include a finding that the product is safe and effective without providing the underlying statistic.

Meaning of Terms Included

In preparing the graphic result above, I searched for a variety of terms that all in some measure implicate accuracy, but at the same time we shouldn't confuse them for the specific term "accuracy." The terms I searched for as well as their common definitions are:

- Accuracy = true positives + true negatives / all results
- Sensitivity = true positives / (true positives + false negatives)
- Specificity = true negatives / (true negatives + false positives)
- Positive Predictive Value = true positives / (true positives + false positives)
- Negative Predictive Value = true negatives / (true negatives + false negatives)

Each of these metrics has different uses in evaluating medical research. An "accuracy" calculation is the most general information and simply measures how many times the test was right out of all times the test was conducted. Regarding the other four metrics:

- Sensitivity, which denotes the proportion of subjects correctly given a positive assignment out of all subjects who are actually positive for the outcome, indicates how well a test can classify subjects who truly have the outcome of interest.

- Specificity, which denotes the proportion of subjects correctly given a negative assignment out of all subjects who are actually negative for the outcome, indicates how well a test can classify subjects who truly do not have the outcome of interest.
- Positive predictive value reflects the proportion of subjects with a positive test result who truly have the outcome of interest.
- Negative predictive value reflects the proportion of subjects with a negative test result who truly do not have the outcome of interest.[1]

As a result, while all five metrics are different, they all are probative of the general concept of accuracy and so I grouped them together for purposes of this study. But, again, 95% specificity has a different meaning than 95% positive predictive value. The graphic result above does not distinguish between the two.

Methodology

From a data science perspective, this is an exercise in natural language processing. I needed to write an algorithm that would extract from tens of thousands of 510(k) summaries the relevant information and only the relevant information. I've been doing this monthly post for a couple of years now, and this was the most labor-intensive study to perform from a technical standpoint. It included a lot of manual work to see if I was getting the right stuff and only the right stuff.

At a high level, here's how I did it:

- I did a significant amount of preprocessing of the data to get the data into a form that it could be reliably searched.
- I then pulled out every single time a percentage was offered, and included several words before and after the statistic.
- Out of that list of snippets, I pulled out all of those snippets that had one of the keywords that I cared about in it.
- Then I had to come up with a myriad of specific rules to get only the percentages that I cared about and not, for example, the confidence interval statistic. What a pain that was. You can imagine the myriad of ways that people express these ideas in text, including the approach of saying "the sensitivity and specificity was 88% and 90% respectively." Bastards.
- Tables proved to be especially hard. The accuracy of my algorithm suffered substantially if the accuracy data was in a table.

I made the decision that I would leave in the truly idiosyncratic stuff that met all of my criteria but still wasn't relevant. I don't think it's much, but I will give you an example. I noticed one summary made the remarkable observation, "Of course no device is 100% accurate." I'm just hoping that not many summaries included such insights. I did read a ton of the output, so I was convinced that such noise was minimal in the ultimate output. It's more likely that I missed relevant output because I didn't have a good way for testing for that, but I started with a pretty wide-open funnel so I'm hopeful that I didn't miss much. I think the sensitivity of my algorithm is good, but the specificity is less well characterized. It was hard to objectively test for the specificity of my algorithm.

You will note also from this methodology that I did not distinguish between clinical and nonclinical testing. I treated all the same.

Interpretation

By a large margin, the most values are in the 95% to 100% performance category. Indeed, 58% of the results are in that category. But that also means that 42% of the results are not. About 15% are in the category between 90% and 95% performance. Add those together, and it means that 73% of the results are above 90% performance.

What about the rest? I should explain that there are 2,335 results presented for about 570 different 510(k)s. That means quite a few of the 510(k)s had multiple accuracy statistics reported in the summary, which is not surprising. Typically, if a 510(k) provides sensitivity, it also provides specificity.

I won't try to quantify this, but I will share that anecdotally I looked at a lot of the outputs and whenever I saw a low number for one result, I typically saw a quite high result for another number. For example, I ran across a submission that had a sensitivity of 70% but a specificity of 98%. Such products might be useful in identifying people negative to the disease or condition at issue, even though the performance is not very good at identifying reliably those who are positive.

There is also, as the methodology above explains, a certain level of noise that I simply couldn't remove but which should be largely ignored.

I think the hump around 70% might also be meaningful. It seems like when FDA looks at sensitivity and specificity, you rarely see numbers below 70%. 70% seems to be a sort of floor to what FDA will consider.

The proportions of the types of devices by clinical context seems relatively stable in each of the different columns. It's not obvious to me that any particular therapeutic area is laxer than others, for example.

Conclusions

I'm leery of offering any particular conclusions because, as I said at the beginning of this post, these results need to be taken with a huge grain of salt. There's a lot of error that I simply couldn't get rid of, and there's built-in bias in the way the data are collected and analyzed. The vast majority of 510(k) summaries do not include performance data, and so in a very real sense the data in the summaries are provided voluntarily by those manufacturers that are pleased with their performance.

However, with those caveats, it does seem as though 95% is the rule of thumb that FDA uses in these accuracy metrics. Having said that, there are plenty of instances where devices are cleared without 95% in at least some accuracy related measure. Often, as I said, it's a matter of a test doing well in one category and then not so well in another, so such tests have a particular clinical function that is not to be confused with ground truth.

As I manually reviewed many of the summaries, I saw products with significant differences, for example, between sensitivity and specificity. To get FDA clearance, the test must do well at something, and at least decently in another category to potentially be considered substantially equivalent to devices already in the market. But in those cases, the labeling needs to be clear about the value of the product and where it comes up short.

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8156826/

National Law Review, Volumess XIII, Number 248