Clinical Chemistry

More Data, Please!

Article:
K. Baggerly.
More Data, Please!
Clin Chem 2013;59:459-461.
http://www.clinchem.org/content/59/3/459.extract

Guest:
Dr. Keith Baggerly is Professor of Bioinformatics and Computational Biology at the MD Anderson Cancer Center in Houston.

Bob Barrett: "More Data, Please!" That's the provocative title of Dr. Keith Baggerly's editorial in the March 2013 issue of *Clinical Chemistry.* He was commenting on an article appearing in the previous month of the journal by Kenneth Witwer who joined us in a podcast earlier this year, regarding the state of data reporting in microRNA studies.

Dr. Baggerly is Professor of Bioinformatics and Computational Biology at the MD Anderson Cancer Center, in Houston. He is our guest in this podcast from *Clinical Chemistry*.

Dr. Baggerly, your editorial starts as a commentary on a paper discussing poor data deposition rates for microRNA data, but you argue that the problems are considerably broader. Well, how broad, and what are the implications of these problems?

Dr. Keith Baggerly: Well, we got involved in a context where we were looking at data from a whole bunch of high throughput biological assays and this involves mass spectrometry, it involves microarrays, it involves microRNA arrays, all of these things that are associated with clinical chemistry et cetera and one of the things that we started noticing is that these are really cool data sets and they allow us to measure tons and tons of things all at once, and say really cool things about the biology.

But at the same time, the very fact that they involve so many things means that we have to keep track of a lot of things and in particular, how all the stuff was processed. And when we started trying to check that in cases where we've seen papers in the literature that said, hey, this is cool, we should use this, hey, this is cool, we should use this. We often found that we weren't able to find the raw data or to completely figure out what had been done with the raw data. And that applied to a whole big group of assays.

Now the reason this is a bit of a problem is that it's not just one or two cases where we can't do it. It's gotten to the point where in some cases folks have now conducted broader surveys looking at the literature, and such is the one by Witwer that looked mRNA rates and found that the deposition rates for the data were less than 50%.

They were ones that have now looked at microarrays and said, okay, let's look at some microarray studies and see what fraction of them based on the data that has been posted, in what fraction can we reproduce the results reported, once we agree on the raw data? And the agreement rates have been things like 2 out of 18 or 6 out of 53 and those rates are really low, particularly when you consider that some of these things are potentially going to be used clinically.

So that's something we need to address.

Bob Barrett: How did you get involved in looking at data quality and why?

Dr. Keith Baggerly: I got involved actually because I am a statistician by training, which means I am a numbers geek, and I work in a cancer center, namely MD Anderson, and we've got a whole bunch of really well-read colleagues who peruse the literature regularly and amongst other things they often find papers that appear to have clinical implications and they get excited by those. And in some cases they will say, wow, this looks really cool; I would love to be able to apply it to treat the patients here better.

But occasionally, some of things that strike them as really cool, involve new modeling of genomic methods et cetera, where they look at it and say, I think this is really cool, but I don't completely understand how it's done. Maybe those guys over in Biostat and Bioinformatics can help. So in many cases this is something where they've come to us, and said this would be really cool if we could get it to work here. Can you help us?

And what we've done in those cases is gone back and taken a look at some of those papers to say, okay, in addition to issues of the clinical implications which we can probably understand, do we understand the fine level details of what's going on? So that's how we got started. What got us more deeply involved was the issue that when we started looking at these in more detail, we started saying, you know there are some pretty basic mistakes in some of these and then we started learning that actually some of these things, where we knew there were mistakes had actually progressed to clinical trials. And we were going, huh, okay,

that's something that needs fixing. And so we tried to do something about that.

Bob Barrett: You have described yourself as working in forensic bioinformatics, and well, you know what the next question is. What exactly is forensic bioinformatics, and why do we need it?

Dr. Keith Baggerly: A friend of mine has said that forensic bioinformatics is a field that should not have to exist, and he is right. Basically, we sort of coined the phrase to allude to the idea of that; in going into the literature, if we can't see how the results were obtained, we will go ahead and say, all right, if we've got the raw data and we've got the reported results, can we infer what must have been done to get from one to the other, regardless of what the published methods say they did?

And in some cases, it's possible to do that. That's part of the detective work that's the forensic nature of things, and that's potentially informative, and in many cases it has uncovered one or two additional steps that the authors didn't realize that they were applying, which has been somewhat disturbing, but it's been informative.

But the problem is that this type of investigation really shouldn't ever be required. The internal process of writing journal articles and reading them is such that the method sections should either spell this out with enough details, such that you shouldn't need to put in a huge amount of time to do this reconstruction, or some of that should be as part of the supplement and things like that. Given that, however, we are in a context where in many cases those aren't adequate, this type of analysis is needed.

And that very statement, the fact that it is needed right now, means that we need to clean up some of the stuff about how journals are publishing papers and what level of data they require before say, this is ready to go.

Bob Barrett: Doctor, can you give us some examples of the types of problems you have encountered?

Dr. Keith Baggerly: Yeah, and some of the problems I will tell you are things that are going to be extremely familiar to anybody in your audience who has worked with Excel. One really cool example has to do with genomics and that actually deals with gene names. A whole bunch of gene names are things like SEPT10 and if you've worked with Excel and you type into SEPT10, do you know what Excel will typically do?

Bob Barrett: I think I do, yes.

Dr. Keith Baggerly: Yeah, so it will helpfully convert that to a date for you. The problem is that all of the scripts that we have that recognize the names and map out to the databases will say, wait a minute, I don't know what to do with the date, so we have this minor problem, and the issue is that with standard gene names, Excel will hopefully convert somewhere between 3% and 4% of the typical gene names in to something else. So we have to be a bit careful about that.

Now that something that's just an idiosyncrasy of the software, but one of the things that can happen is also that it's possible when you're working with large datasets to introduce subtle errors and possibly not notice it. For example, if you've got a small table's worth of data, and it is say four columns and five rows, if you make a mistake and you offset one of the columns by one row, it's really visible, you look at it and you say, whoops, I screwed up, and you just fix it.

If you make that same type of mistake with a table that has 100,000 rows, it's often not as easy to spot, and the problem is if you are working with 100,000 things or 100,000 genes or something like that, it's often the case that if you are saying, well, this set of a hundred is probably important, I've yet to be convinced that anybody, be it a biologist or bioinformatician or whatever, really has a really great understanding of what a hundred genes are doing all at the same time.

And what this has led to in some cases, is the context where at least in one context, people looked at a gene study where they read in a table of gene expression values and gene names, and they said, we've come up with the following list of a hundred genes that is important for predicting whether or not this patient will respond to chemotherapy and this makes sense, because the following genes are here and these genes are associated with chemo response.

And that would have been really cool, except, the software they were using required two inputs, the names and the numbers and the thing is for the numbers, they said we want a header row telling us what this is and for the names they said we don't want a header row at all. And the problem is people cut-and-pasted from Excel. So they gave it a list of names with a header row, so all of the genes were offset by one, in alphabetical order from the ones thought they were working with.

So this list of a hundred genes which was going to predict chemo sensitivity, *which made sense*, because it involved certain genes actually didn't involve the genes that they thought it did at all.

So it's something where our intuition can trick us and that's something where it's a simple mistake, it's an off by one and you can see how people might do this. This is not intentional in any means. But it's something where that type of thing can have broader implications. A more dangerous type of implication has to do with the fact that often times people if they are just writing down things that they want to code for one or two levels, they will just code it as 0 or 1 or something like that, and that's perfectly fine as long as everybody agrees what that is.

Problem that happened at least once is a scenario where again looking at genomic data, a bunch of patient samples were coded as 0 and 1, in this case however, 0 and 1 were supposed to indicate whether the patients were responsive or resistant to therapy, and as they went across a few tables like this, for some of the tables they got the numbers reversed and then they were making predictions based on those tables.

Unfortunately, because those labels were reversed what that would amount to, is if you are making a prediction that this patient is likely to be sensitive to this drug, we should give it to him. Then if your method works, you're actually going to wind up administering the drug you have just to the patients who are most resistant to it, which is a bad idea. But that's a subtype of simple mistake and the thing is that's actually the type of mistake that's by far the most common. People aren't trying to make mistakes; they make easy-to-fix mistakes. There is however the problem that at present in the scientific literature; I alluded to the fact that in many cases the documentation isn't keeping pace. We don't have the full description of what was done.

So if you take a simple mistake and then you don't document fully what was done; that simple mistake now becomes hidden and harder to find and to fix, and that's where things can get bad. That issue that I just alluded to of; here is the thing where we have some patients labeled 0, some patients labeled 1, that's actually a case where that dataset made it into guiding therapy in clinical trials and that disturbed us.

Bob Barrett: Are all these problems fixable, are you optimistic about that?

Dr. Keith Baggerly: Yes I am. Although -- so I get to go often, I talk about some of these problems that exist and how they are widespread and things like that and that makes me sound something of a downer, or something by saying, here is all these things that are out there--but one of the things that needs to be realized is that while these problems are at present more widespread than I would like, there is nothing inherently

unfixable about them. And indeed the very nature of the things, the things that these are simple mix-ups or whatever, say that if we can just train ourselves to keep track of these things, then they actually should be fairly easily fixable.

And this issue of training ourselves to keep track of things, believe it or not, this is not a new problem. Other people have run into this and solved it before and in particular what we are now seeing is it's coming up as a problem in, and in particular, in molecular biology because you are having a clash of a system of working which has worked pretty well for laboratory experiments, conflicting with a massive volume of data, which it just hasn't had to deal with before. And the tools for keeping track of many of these records et cetera have actually already been developed in many areas, in particular things like software code development.

So there's a bunch of tools in the Open Source Community for keeping track and doing all these data checks and saying, have you done this correctly and can we trace exactly what was done? And I will happily admit that in our bioinformatics group, we are happy to steal these ideas and use them here. We are not proud; as long as it will make the job easier and allow us to do our jobs better.

So the fact that the tools exist is pretty nice. It's also something where because this problem of lack of reproducibility is now being recognized as an issue, just in the past few years you have several groups begin to develop new tools to make it easier to write code and write results in such a way that the full documentation will be there at the end of the day. And one particular tool that we've looked at in some detail is actually called Markdown and we may have played a bit with HTML which is a Markup Language.

Markdown is trying to make something, like a Markup Language, but they want to make it simpler. And the way that they have done it is the authors of Markdown have now looked at how people have written email for the past several years, and they have said, people have gotten used to doing these little smiley faces and if they want to make things important, they will stick an underscore on either end. And they said you know what, people who found good empirical shortcuts that emphasize the points that they want to make, let's go ahead and use those as the basic syntax for what we want to do.

And that actually is something that has turned out to make writing code in that syntax actually pretty easy. And it actually is something that allows us to track and make these things more reproducible and better. And this is something

where, okay, I am a senior member in our department here and this is something where I sat down with one or two of the other faculty here and we looked at the stuff and we said, wow! This is really cool!

You know what; our department is not a democracy. So we got together with some of our analysts, the folks who are working with us to produce reports and we said, guess what, we are going to write reports, this way, going forward. And that has actually been fairly easy to implement and we have been very happy with the results.

Bob Barrett: There's certainly an awful lot to chew on here. How can the readers of *Clinical Chemistry* help in the process and where can they go for more information?

Dr. Keith Baggerly: Well, in terms of what they can do to help with the process and things like that, actually the American Association for Clinical Chemistry has been looking at the issue of reproducibility and there has been some discussion of putting together a position statement saying, what are the goals of the Association? This is motivated in part, by the fact that I mentioned a moment ago that some of these problems have actually made it into clinical trials.

The fact that things got that far meant that the Institute of Medicine, the IOM, got together and put together a report saying, what types of evidence should be required before these types of complex tools are used to guide therapy, and they issued a report on that in 2012 in March saying, here are the guidelines, here are the things that we think should be in place. And there is a committee of the AACC, American Association for Clinical Chemistry, that has reviewed that and is looking at putting together a position statement on that.

And they have been trying to formulate that for -- and I don't know precisely when they are going to issue it, but it's something where they are saying, here are the types of information we've got to have to put down.

Now, as a shortcut to that, while the precise details are still being formulated there, I will point out that, well, we actually wrote an editorial for *Clinical Chemistry* in 2011, saying what types of data should accompany OMICS publications, so there is that. There is some of the stuff that Witwer is now recommending for microRNA papers and his recommendations are pretty good.

And if you really want to know more about how these issues are coming up and how they've actually turned out to impact clinical care, well, the next thing I am going to do is I am going to point you to Google, and I will say go out

Google “Baggerly and YouTube,” and watch the video you find there [http://youtu.be/7gYIs7uYbMo] that says, here is reproducible research, here's how bad things can get and here are some of the gory details of some precise examples of what actually went wrong, and how things got fixed. And if you want to do more, there are things like, now, a whole bunch of reproducible research focus groups on the web and blogs et cetera, and it's actually sort of fun and it's neat to see in some sense, in my view, a stronger level of discourse, in the scientific literature, begin to emerge as people say, you know what, we're going to do it better. And that’s what I got.

Bob Barrett: Dr. Keith Baggerly is Professor of Bioinformatics and Computational Biology at the MD Anderson Cancer Center in Houston. He has been our guest in this podcast from *Clinical Chemistry*.

I'm Bob Barrett, thanks for listening!