Host: This is the podcast from *Clinical Chemistry*. I am Bob Barrett.

Recently, Keith Baggerly and Kevin Coombes, two biostatisticians at the MD Anderson Cancer Center examined data published elsewhere, aimed at predicting the response of a tumor to chemotherapy. Their work published in the *Annals of Applied Statistics* reanalyzed a set of experiments and uncovered a series of errors, including mislabeling and cases of where the gene probe identifiers were mismatched with the names of genes.

The September issue of *Clinical Chemistry* published a perspective article by Dr. Stephen Master on this emerging discipline of forensic bioinformatics, which is an effort to reconstruct and validate analytical results that have been reported in the literature.

Dr. Master is an Assistant Professor of Pathology and Laboratory Medicine at the University of Pennsylvania, and Director of the Endocrinology Laboratory at the Hospital of the University of the Pennsylvania. He continues the discussion of the report on gene signatures for cancer therapy in this podcast.

Could you tell us the background of this work, Dr. Master?

Dr. Stephen Master: Well, this work was initially spearheaded by a group from Duke University who had an interesting and potentially very powerful idea. They were interested in looking at gene expression data from a set of previously characterized cancer cell lines, and the reason that if you compared cell lines that were sensitive to a certain drug, and other cell lines that were resistant, then you could identify the gene signature that correlated with this sensitivity.

Then the idea is that you could take that gene signature and examine primary tumors from patients, and the hope would be that you could use this as molecular diagnostic to predict whether an individual patient's tumor would be likely to respond to a certain kind of chemotherapy based on the original cell line signature. So, the reported results of these experiments looked quite good, and the work was published in a series of papers over the past four years or so.

Because it looked like a successful and valuable example of personalized medicine, it led to some real clinical excitement, including an MD Anderson where Baggerly and Coombes were working. So, Baggerly and Coombes wanted to understand this work from the Duke group in detail, so that they could apply it at their own institution. They decided to start by reconstructing from the raw data, exactly how the final results had been obtained.

Host: So, what did Baggerly and Coombes find through their reanalysis?

Dr. Stephen Master: Well, I think the first thing to point out is that this type of analysis can actually be quite difficult. So, of course, you need the raw data to begin with, but the actual path from those raw data to the published results is not always entirely clear. There are a number of particular complexities to the bioinformatics of processing these kinds of data. There is often also limited space devoted to describing this in the Method section of a publication.

So, that's why there was a need for what Baggerly and Coombes have called Forensics Bioinformatics, which is really just trying to reconstruct this process to fill in the gaps to show the exact steps that led from the primary data to the final analysis. Now, what was important in this case is that as Baggerly and Coombes looked in detail at this data, they began to uncover a whole list of problems.

As you mentioned at the outset, these included problems like mislabeled samples, duplicated samples, incorrect gene list, and even actually inclusion of genes that apparently hadn't been measured on the microarray at all. As they looked further, they found that many of these problems recurred in multiple papers and through multiple analysis. So, their conclusion was that these types of errors occur fairly commonly. It's actually fairly easy to make some of these mistakes when working with large data sets.

The case of the gene list is a great example of this, because it appears that this was all simply from not accounting for header line in the data file, so each gene list that was actually off by one from what it should have been. Of course, this is easy to do when you're reading in a data file, but those of us who work with spreadsheets on a regular basis recognize that this can also happen, anytime a data set is being manipulated by cutting and pasting, by hand, for example. So, it really underscores how a simple mistake can throw things into real confusion.

Host: What do you see as the most important lesson from this controversy?

Dr. Stephen Master: Well, I think there are at least two very important lessons. The first is, I think that we would need to recognize that it's very hard to detect these errors when you're just looking at final results, from a complex, genomic or proteomic data set. I think this means that we as a scientific community need to get much better at making sure that each step along the way is adequately documented, so that it becomes easier to pick up a mistake at step 2, let's say, that may be

difficult to see by step 10, but which may completely change the final result.

Given the fact that we have different groups of investigators using different algorithms, different software packages, different parameters for those packages, this can be difficult. But I think that this is the most important aspect of reducing these types of errors.

Now, the second lesson, I think, is that it's critical to pay attention to data management. Quite often, we think it's a difficult and expensive part of doing a genomics or proteomics experiment is acquiring the data, using the DNA microarray or whatever platform you happen to have. Of course, it's true that if you don't do that part well, whether you've bad lab practice, you don't randomize your samples or whatever, you certainly can ruin your experiment.

But what Baggerly and Coombes are showing, I think, is that the data management piece can be every bit as important for obtaining good results. I think that this is a lesson that applies at every level, from an individual laboratory on up through an entire health system.

Host: Well, could anything else have been done to prevent this?

Dr. Stephen Master: Well, one of the real question that I supposed comes to mind, to mind of many people is, with all these problems, what did the papers get the literature in the first place?

(00:05:01)

One problem, of course, is that no reviewer would have time to trace the data in the careful way that Baggerly and Coombes have done. This may suggest that journals will have to give detailed thought to what they will consider adequate documentation for bioinformatics.

Actually, one of the great things, I think, about this incident is that it's even rise to groups that are working actively on this very problem, deciding how much information is required in order to make it possible to reproduce bioinformatics results without resorting to hundreds and hundreds of hours of forensic bioinformatics.

Host: So, does this problem just affect labs that are working with microarray data?

Dr. Stephen Master: Well, actually, no, I would argue. To me, one of the most interesting things about the work of Baggerly and Coombes, is that it points out that any time you're working with large data sets, whether these are genomic data, proteomic data, or perhaps even large batch data sets that may be handled

by translational research lab, it can be very difficult to spot errors. Any of us who work with large data sets are susceptible, and this once again really underscores that need that I talked about to invest very carefully in ways to manage this data.

Host: This story has recently made its way into the mainstream media, in your opinion, what are the implications of the latest developments?

Dr. Stephen Master: Well, the largest story surrounding this case has expanded substantially. When Baggerly and Coombes published their analysis late last year, Duke temporarily halted clinical trials that had included gene expression signatures as one component of the research. After a confidential review at Duke, the studies were restarted, although, perhaps without full clarity, as to how the scientific issues had actually been addressed.

What happened next though is that there have been allegations that one of the principal investigators had actually falsely claimed award, such a Rhodes Scholarship on a CV. This was reported by a number of sources ranging from the *Cancer Letter* to the *New York Times*. So, the clinical trials have once again been closed to enrollment. Now, I think that this new development may put a different spin on some of the irregularities in the data management problem, but I also very much think that it's important that the scientific lessons not be lost in these new problems.

The fact is, whether or not this new scandal had emerged, it is still the case that there were a number of significant mistakes in data management at the scientific level, and these mistakes were very difficult to spot without an extensive foray into forensic bioinformatics.

Host: Well, it's obvious, this topic has significance with the research community. Will it also affect clinical and medical laboratories?

Dr. Stephen Master: That is a great question. I think the answer is yes for at least two reasons. First of all, all of us who spent time in the clinical laboratory community are well aware of the need for new clinically relevant biomarkers. Now, a number of reasons why these markers have been slow and coming, but I think the lesson of Baggerly and Coombes work is that one of these is errors of data management. If diagnostic developers, whether they're an academia or whether they're an industry, don't pay very careful attention to this kind of an issue and pay attention to good data management, very clear documented pipelines for analysis, then the types of errors found by Baggerly and Coombes can easily creep in, as they have demonstrated.

This slows development. It could have harmful side effects of discrediting the field and at the end of the day, it hurts us as we attempt to improve patient care. So, I think that this is the first way in which this is very relevant to the clinical lab. Secondly, I do think that this may provide another opportunity for the clinical laboratory to be of some help. Everyone involved in running an accredited clinical lab, knows the kind of validation of data transfer that needs to occur in order to deliver quality care. I hope that this level of rigor will allow us as a clinical laboratory community to also contribute positively to the conversation about data management and the research lab.

Host: Dr. Stephen Master is an Assistant Professor of Pathology and Laboratory Medicine at the University of Pennsylvania and Director of the Endocrinology Laboratory at the Hospital of the University of Pennsylvania. He has been our guest in this podcast from *Clinical Chemistry*.

I am Bob Barrett. Thanks for listening.

Total Duration: 9 Minutes.