



**Article:**

K. Witwer.

*Data Submission and Quality in Microarray-Based MicroRNA Profiling*

Clin Chem 2013;59:392-400.

<http://www.clinchem.org/content/59/2/392.abstract>

**Guest:**

Dr. Kenneth Witwer is Assistant Professor in the Department of Molecular and Comparative Pathobiology at Johns Hopkins University in Baltimore.

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

Public sharing of scientific data has assumed greater importance as increasingly massive amounts of data are being generated by various sequencing procedures. Transparency is necessary for not only confirmation and validation, but also when extracting maximal value from these large data sets. To address this issue, guidelines on the minimum information that can be published from microarray experiments have been promulgated and used in the peer review process in the acceptance of new papers.

In the February 2013 issue of *Clinical Chemistry*, a paper by Kenneth Witwer examined just how well these criteria are being followed. Dr. Witwer is Assistant Professor in the Department of Molecular and Comparative Pathobiology at Johns Hopkins University in Baltimore. He is our guest in this podcast.

Dr. Witwer, in your review, you evaluated studies of a particular type of nucleic acid, microRNAs, and how well those studies follow the guidelines of the journals in which they're published. First of all, what are microRNAs and why do they matter?

Dr. Kenneth Witwer: Well, microRNAs are encoded in our genome, but they have effect on the genome that go far beyond their sequence. I often think of our DNA as something like a recipe, so that if a dish that I had last night had been made from a particular recipe, that could have been made at any restaurant, anywhere in the world, but at each of those restaurants it would have been slightly different.

So the sequence of the DNA is certainly important for things like development of disease, but it's not the only factor. So microRNAs are more like a chef, I would say, at one of these restaurants, and the chef would of course add her own unique flare and her ideas and practices to the mix.

So, a DNA is like the recipe, and microRNAs are something else that we call the epigenome. So I believe that you understand how a disease might develop, it's often not enough just to look at DNA sequences. And there is another aspect to microRNAs that's also quite interesting, potentially for clinical practice and that's their application as biomarkers.

So because of the size of microRNAs, it's very small and they can be packaged into vesicles and other carriers that can carry these microRNAs out by the cell, that is to say that if a particular group of microRNAs is made by tumors cells, we might be able to find these RNAs in blood or in other bio-fluids like urine, or saliva or even cerebrospinal fluid, and we could diagnose cancer or other diseases without having to take a biopsy for example.

And I think that this possibility has attracted a lot of interest in the research community, and investigators are generating large data sets in which they measure hundreds of, or even thousands, of small RNAs in bio-fluids and also in tissue.

Bob Barrett: So, microRNAs may have tremendous implications for new clinical tests and we'd hope that reporting standards are followed to make sure that the best tests are developed as quickly as possible. What did your review of published papers reveal?

Dr. Kenneth Witwer: Well, just to give a little bit of background, the scientific community decided over a decade ago, that large data sets that are collected with microarrays, so these are often glass slides that have oligonucleotide printed onto them and that can be used to determine gene expression, or in the case of this study microRNA levels; that these datasets should be reported publicly when a researcher wishes to publish results.

So the rationale for this was that these experiments are very expensive, they are very complicated and that's true on the analysis side just as much as it is on the wet lab experimental side. So that initial initiative was known as MIAME or the Minimum Information on a Microarray Experiment, and today there are analogous initiatives or guidelines that have been taken up by other fields, such as MIAPE, that's for proteomics experiments and something called MIQE, that's for quantitative PCR.

And to comply with these standards including MIAME, I as a researcher would be expected to upload my data to the public database and this would include both my raw data and then the data that I have processed and normalized and analyzed somehow myself, as well as a very detailed

explanation of all of my methods. And the goal of this is to present a data set that another researcher anywhere else in the world could take, could follow my methods, and then replicate my results to confirm that everything was okay.

So I started out by looking at over 120 articles that contained microRNA microarray data, and that were published between 2011 and 2012. I focused on journals that published the largest number of these articles. And what I found was that only about 40% of studies included full data submission, and only about a quarter of them were fully compliant with the reporting standards known as MIAME, and this means that the various aspects of the analysis, of the processing, were not included in these reports.

I think that even, perhaps even more worrisome than that was, I also found that when data sets were analyzed that had been uploaded perhaps after the publication process--so after say a journal discovered that their guidelines had not been followed--it turned out that a lot of those data sets were not fully supporting the conclusions that were made in the papers.

Bob Barrett: So a part of what you examined is the public availability of large data sets. Scientists that receive NIH funding, have an obligation to make work available to the public and the NIH and its counterparts in other parts of the world have established websites where researchers can store their large data sets, free of charge.

Now what you are finding is that many groups are not taking advantage of this opportunity?

Dr. Kenneth Witwer: That's right Bob. So even despite these certain mandates many studies do not include database submissions.

Bob Barrett: Is the issue of reporting specific to microRNA research, or other types of studies also affected, and what ultimately is the harm when someone doesn't submit scientific data to these databases?

Dr. Kenneth Witwer: Well, to the first part of your question, no, I don't think that this is specific to microRNAs. Now microRNA research is relatively new compared with say transcriptomics research that really had a solid start well before the last decade.

So I think that there are lot of researchers who are coming into the microRNA research field who do not have a lot of experience with large data sets, so that could be contributing to any possible differences between microRNA and other forms of research. But no, I don't think that this is specific to microRNAs.

To the second part of your question, the harm here is manifold, because data can be lost. And in some cases when journals have followed up with authors and asked them to submit data that were supposed to have been sent in with the publication, the authors were unable to find that data, and of course, this is a tremendous loss, not just to the research community, but to anyone who is concerned about public health, because in many cases, these experiments have involved very precious patient samples that cannot easily be collected again and analyzed again.

So these public databases are a very important resource for the life sciences, and are important for data security. So another aspect to this is that when data are not available, the reviewers of a publication don't have the chance to properly review a publication. And then at another level of colleagues or readers of the publication can't confirm results before starting their own studies, and because of all of the time and the resource expense that's involved in replicating studies or building on existing studies, this can be quite wasteful.

And then finally, I would also like to note that whenever I publish a data set, there is probably a lot in there that I'm not able to extract myself. I may not have all the analysis tools or maybe I don't have all the ideas that others might have. So someone else could come to my data set with a novel idea and extract something completely new from it that I would not have done. And I think that it's important to have those data sets available so that we can maximize the impact of the results that we're achieving.

**Bob Barrett:** Why would researchers not want to take advantage of public databases, to not only preserve, but promote their work?

**Dr. Kenneth Witwer:** Well, that's an interesting question, and several year ago, in 2011, Wicherts and colleagues reported in PLOS ONE that an unwillingness to submit data is associated with poor quality of a study, which I thought was a very interesting hypothesis and this study was written from a completely different field to my own.

It was the field of psychiatry and they were looking at data that were available for published studies. And to rephrase their hypothesis and their conclusions, I would say that if I'm worried that my data or my analysis are somehow weak, or don't fully support my conclusions, I probably will tend not to want to let others look at it too closely.

**Bob Barrett:** Do you think that anyone is trying to mislead scientific colleagues or the public, by withholding data sets that they know be substandard or flawed?

Dr. Kenneth Witwer: No, I don't think that's really the main issue here. It could be that there's the subconscious knowledge that there is something that's not quite right about data sets, and so, sometimes people don't wish to upload them for those reasons. But that's probably not the main reason that data sets are withheld.

I think that in many cases, it's a simple matter of an unawareness of the data reporting requirements. But then there's also this issue of researchers who feel that they own their own data, and that's an understandable assumption, but it's also incorrect. I think back to a meeting that I attended several months ago, where one of my very respected colleagues got up and said that he didn't wish to submit data because he didn't want to have a labless lab going through his data and publishing something, based upon what he had found.

And while I respect this researcher, I also strongly disagree with a statement like that, because I believe as you alluded to earlier, any publicly funded researcher has the obligation to share data with the community, and to make that data available to anyone who asks for it, and the journals have made this easier, they've facilitated this process by requiring those data submissions for publication.

So the assumption that I own my own data is simply incorrect.

Bob Barrett: Well, finally doctor, what can be done to ensure that important research results are stored and preserved and available in a way that will maximize impact on public health?

Dr. Kenneth Witwer: Part of the answer to that question, I believe, resides with the journals. The journals are, for better or worse, the gatekeepers of the scientific community. And so the journals, and by journal I mean the editors, the reviewers, the staff, all working together, should take steps to ensure that all of the necessary data are available before the review process even occurs.

And this will allow scientific reviewers then to take a look at those data, if necessary to even analyze the data or do spot analyses to confirm that the data sets have full integrity and can be used to support the conclusions that are made in the paper. And I think that researchers also have an obligation to check with journals about those reporting requirements and even to upload the data without prompting.

I believe that if we all develop the mindsets that our data are part of a unique opportunity that we all have to

contribute to public health, then we will start to see that there is a vital importance of making these data available to as many people as possible.

And ultimately I also believe that this is good for the researcher. So instead of seeing the data upload process as a way of losing control of my data, it's actually a way of making sure that my data are available to as many people as possible, and will have the greatest impact possible.

Bob Barrett:

Dr. Kenneth Witwer is Assistant Professor in the Department of Molecular and Comparative Pathobiology at Johns Hopkins University in Baltimore. He has been our guest in this podcast from *Clinical Chemistry*.

I'm Bob Barrett, thanks for listening!