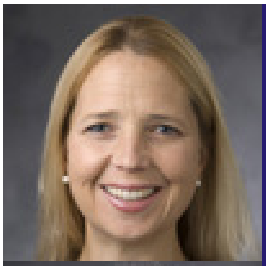

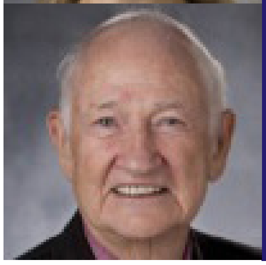

AN INTERVIEW WITH

DR. RACHEL RICHESSON & DR. ED HAMMOND

Co-chairs of the NIH Collaboratory Phenotypes, Data Standards, and Data Quality Core

Interviewed by Liz Wing, MA, Coordinating Center Staff Writer

At the May 2017 NIH Collaboratory Steering Committee meeting in Bethesda, MD, we sat down with Rachel Richesson and Ed Hammond and asked them to reflect on the first 5 years of their Core as well as on the challenges ahead.

How would you describe the first 5 years of the Phenotypes, Data Standards, and Data Quality Core?

Richesson: It's hard to imagine it's been 5 years, and though we're often referred to as the Phenotypes core, actually our scope is quite large. Early on we realized that data quality and data standards were also under our charge, but certainly phenotypes has been our main focus in these first years.

Hammond: In the first year of the Collaboratory, I was excited about thinking in terms of phenotypes and felt that the Demonstration Projects were a wonderful opportunity to explore definitive data elements and data issues. We found that our concept of phenotypes was not as simple as cohort identification because there are so many factors in a trial that come into play. The projects were just starting, and we all struggled to get clarity on how the study teams were planning to collect their cohort data.

Richesson: That first year we were getting to know the study teams. Their pragmatic trials were diverse, as were the diseases being studied and the partner healthcare organizations. We also had diverse membership on the Phenotypes Core—clinicians, informaticians, epidemiologists, and health system leaders. Our plan was to ask the study teams to give us their phenotype logic so we could create a dictionary to share with others doing pragmatic trials. We wanted to standardize how the teams were capturing their patient populations, but this proved challenging because of the multiple disease areas and the trial-specific components.

Hammond: Rob Califf in his keynote address at this meeting raised the notion of developing "meta knowledge." I think phenotypes may well be meta knowledge because they go beyond just cohort identification. For example, you may have a number of different data elements that are part of a phenotype definition. If you meet 3 out of 5, or 1 out of 5, what does that really mean? I think that phenotypes may be a way of discriminating between different manifestations of the same disease. Phenotypes are a way of repackaging a lot of things other than cohort identification. I think that's what the future is going to be.

Richesson: And in the future, rather than top-down development of phenotypes, there will be other approaches that emphasize how the data speak to these definitions as well as more use of machine learning and probabilistic logic. It will be interesting to see how we standardize these approaches and share this information going forward.

> "One of the major contributions of this Core is highlighting the importance of data quality."
>
> *– Hammond*

NIH Collaboratory
Health Care Systems Research Collaboratory

What accomplishments of your Core are you most proud of?

Richesson: I'm most proud of the data quality assessment and reporting guidelines that Meredith Zozus of our Core started during the first couple years. They provide guidance for how to look at data from different places to identify if you're indeed collecting data for the same populations. That guidance emphasizes looking at the provenance of data and trying to understand the workflows so we know where that data came from. This product of our Core has been a good resource for the Collaboratory.

Hammond: The Core's work has influenced my research a great deal. For example, I'm interested in looking at temporal phenotypes and what's being tracked over time. There may be people who match a phenotype at an instant in time, but how do they change over time from that single point? This could have significant value in terms of understanding different diseases, pathways, and timing. I think phenotypes are still a good way of representing the change of symptoms and the course of a disease within individuals.

> "We've built a community in our Core that represents a diverse group of scientists and clinicians showing the many ways to look at data challenges."
>
> *– Richesson*

What do you see as the biggest impact of your Core to date?

Richesson: We've provided guidance documents and contributed to the PCT Reporting Template on the Collaboratory's website, which clearly states that researchers who are designing and reporting trials need to be explicit about the data they're collecting and the populations they're identifying. We've also contributed a couple chapters to the Living Textbook that give a high-level view of issues with phenotypes and use of electronic health record (EHR) data. More importantly, we've built a community in our Core that represents a diverse group of scientists and clinicians showing the many ways to look at data challenges.

Hammond: Data quality is becoming increasingly important. I'm interested in ensuring that quality is part of the collection of the data, and we'll find new ways of capturing data. Data quality certainly plays a part in the value of the EHR as a source of data for pragmatic clinical trials. Quality is completeness and consistency—and the ability to answer the research questions correctly. One of the major contributions of this Core is highlighting the importance of data quality.

What work is important to tackle going forward?

Richesson: We've done good work with phenotypes and laying the foundation for data quality, but now we need to address data standards and identify smarter ways to capture data at the clinical settings. I'd like to see us find a message that we can advocate and contribute to the development of regulations and standards in how clinical data are collected to support pragmatic research and learning healthcare systems.

Hammond: The focus of big data is changing rapidly. My hope is that, instead of saying EHRs are not good enough to do observational clinical trials, we fix the problems with EHRs. It's important that both clinical data and research data are accurate. We're also beginning to see other kinds of data, such as behavioral, social, environmental, economic, and genomic data. But we don't yet know how clinicians will use these. Temporal phenotypes may be a way of taking advantage of different types of data by including them in the clinical equation. For example, we could build environmental data into the phenotype itself. Then we could predict outcomes as well as the proper interventions. This is an area with depths still to be discovered.