Educational Perspective

Why is the Kern Institute's Medical Education Data Scientist Smiling? *She is building a lab to put it all together!*

by Tavinder K. Ark, PhD

Call me Isthmus: a person who connects cognition, statistics, psychometrics, data and programming with adventurous fun. I love organizing, analyzing, interpreting and visualizing data that clarifies how and what our learners learn. There are a number of fancy titles for the sort of person who can and likes to do all this. I could appropriately be called a data or educational measurement scientist or more colloquially a data Jedi. While I am aware that some people see all this as a "black box" of fancy data footwork, if you give me some time, I am committed to demystifying the magic so that everyone understands how data science is a critical part of our medical education transformation work and why the Kern Institute is building a medical education data science laboratory.

What does it mean to love this work? It means I wake up at 1:00 a.m. with palpitations wondering if I have coded the response options on a survey from strongly disagree to strongly agree with the correct numbers (1,2,3,4) in my analyses. I spend countless hours engaging in good 'data hygiene' practices – writing computer code to "get data out" of data collection software – such as Qualtrics, Survey Monkey, ExamSoft or RedCap – into easily digestible formats (a little fancier than Excel-like software) so that I can use open source data analysis software such as R to make sense of it all. This type of highly-detailed work used to take many people weeks and be riddled with errors. I help create automated, frictionless "pipelines" that offer a generalizable, reliable data in seconds. Sounds great right? But why do it?

My goal is to create a data repository and visualization system that makes it not only possible, but easy and fun for learners and teachers to master their crafts by gathering high quality feedback to guide their work.

- What if, as the teacher, you realize that an hour after a physiology session – where every student understood the factual knowledge – half the group could not apply those facts to a clinical case? How might you change what you do next time to get everyone up to speed?
- What if, as a student you could walk into a clinical foundations seminar reassured that you knew 80% of the material and could focus your time on mastering the part you don't know?

I can imagine making this possible, but it will take some significant work.

One key element of a medical education data science lab is data linkage. If we could follow the same group of individuals over time and understand their individual learning trajectories as they work their way through our curriculum across courses and clerkships, we could implement programs that maximize learning and makes teaching and the role of teachers much more interesting. By analyzing the data in interesting and innovative ways, we can better understand our students, creating a subtle understanding of how to help those who are struggling with some of the material, and allow others to move through faster. We are working to enable doing this within an ethical framework so that privacy is protected, and the benefits to learning and professional development far outweigh the risks.

Loving this work also means I feel defeated when I cannot find significant differences in the measured learning of students when they complete a curriculum designed to improve learning outcomes, or improvement in the health and well-being of students after an intervention that was designed to make them feel better. My eyebrow is raised, and my brain is churning when a longstanding theory about the relationships between constructs – such as medical competencies in medical education – are not confirmed empirically. This is the case, at times, with AAMCs Core Entrustable Professional Activities (EPAs) for Entering Residency. But, that is why we call this science. Figuring out

why it did not work is where the fun often begins in what I call the “dance” with data. It really is symphony of analyses, with me as a conductor trying to find where both the noise and signal are coming from in the orchestra of data.

In this process, I must be innovative, utilizing a full range of lenses through which to view data. In all this, I need to work with my best friend, the computer, and tell it how to run accurate representations and in-depth visualizations, appropriately slice the data and conduct statistical tests that separate the signal from the random noise. All of this results in interesting and previously undiscovered connections between the constructs we have measured.

I spend lots of time designing studies that isolate the variables of interest (e.g. what students learned) from those variables that just confound the picture, designing new measurements and plans to validate new tools that actually measure what we hope they will. This involves contextualizing the current objectives of curricula or interventions in the literature and identifying existing surveys or measures that help us do our work. Once we have collected data, I spend a great deal of time just looking at the data in a variety of ways, graphing, modeling and testing to see what our data can tell us about what we “think” we know and teasing out relationships among the data that were not anticipated.

This combination of theory and data driven analyses involves hitting the books as much as running analyses. I spend a lot of time in the descriptive landscape, slicing, dicing, and re-slicing using a lot of data visualization techniques. This is where the data comes to life showing nuances that are otherwise missed when we roll things up to just a “mean” or significance testing of two or more numbers.

Once I am convinced we have found something “truthful” and meaningful in the data, I spend time designing ways for our learners and teachers to see data in an aggregated, digestible, graphical forms so they can implement evidence-based and data-driven changes in their everyday behavior, professional development, and life-long learning. When done well, this is a powerful source of feedback for continuous improvement of learning and teaching.

Finally, after we finish any study or curriculum, there is still a lot we do not know regarding why our learners scored a particular value on a test or answered a series of survey questions the way they did. In statistics, we call this the error term and what most understand as "noise." The noise is where I would say the "pudding" is. It is all the things left that we have not yet measured, have not understood, or have misunderstood. Some of this error is just meaningless and random, but some is not. Spending time with the "systematic" error helps us understand things we do not yet understand. This is where data scientists can really have fun because it requires careful unpacking and pushing theories beyond their bounds, coming up with innovative research designs to isolate variables of interest, and evolving our understanding of statistical analyses that can help us tease apart noise from signal. It is a deeper dive into what we do not know and into what is missing. In the sandbox of data, bring your calculator, but definitely do not forget the ruler.

Tavinder K. Ark, PhD is a data scientist and Assistant Professor in the Robert D. and Patricia E. Kern Institute for the Transformation of Medical Education.

Return to *Transformational Times* by selecting the browser tab to the left at the top of your screen, or by using your browser's back arrow.