Podcast Interview: Atul Butte

PNAS: Welcome to Science Sessions. I'm Leigh Cooper. Nowadays, many journals and grant organizations require researchers to upload their experimental data to data repositories. A biomedical researcher interested in data on RNA, DNA, specific diseases, or even drug trials can likely find an online repository brimming with relevant records. National Academy of Medicine member Atul Butte of the University of California, San Francisco has long recognized the value in repurposing these mounds of data. For example, previously, Butte and his colleagues explored the genetic underpinnings of hepatocellular carcinoma, a type of liver cancer, from publicly available gene expression data, and searched for currently available drugs that target those molecular signatures. That effort ultimately revealed that a deworming pill could be modified to act as a promising drug candidate against liver cancer. I talked with Butte at the World Conference of Science Journalists in San Francisco last October, about the potential obstacles to using big data for biomedical research. He says that finding the data is the easy part. The hard part is asking the right questions.

Butte: What we are starving for are the questions, because the data itself doesn't ask an interesting question. It still depends on an individual to come see what's there and ask the question. And, by the way, we still call people who ask questions scientists. So these are not just computer programmers here. These are people who can see what's doable and then ask that right question that generates that hypothesis that just happens to be askable and answerable using data.

PNAS: Some researchers worry that, by allowing access to their data, they might get scooped and miss out on a publication.

Butte: So when we're doing this kind of science, we're almost always asking questions that the originators of the data never thought about asking. And maybe they didn't have the right context, or maybe they didn't realize their data set could be paired with another one to enable a new question. So what that means is that we're almost never stepping on people's toes here. In general, the few papers out there studying open access have shown that your citation count tends to go up if you actually release some of your data.

PNAS: When choosing studies for data mining, Butte recommends picking newer studies, as experimental and data collection methods generally improve over time. And, Butte says, the amount of data in these repositories is doubling roughly every three years. But what is the quality of this data?

Butte: People argue garbage in, garbage out, but the best researchers get the money to generate this data. Don't wait for perfection here, right? Voltaire has the famous quote. Perfection is the enemy of the good. The data is good enough now for a lot of things. Try to figure out what can it be used for in its current form today.

PNAS: Butte says that he doesn't trust any single dataset he would get from a repository. But, then again, he doesn't have to settle for just one dataset.

Butte: I prefer to get 10 datasets or 100 datasets or maybe even 1,000 datasets. Maybe different models, maybe in certain areas people fight over biopsies are the right way to study it or cell lines are the right way to study it or mouse models are the right way to study it. I say take all three and try to figure out what's in common. What are they all seeing collectively together?

PNAS: Butte says that data may not be available for some rare diseases, but that many disorders have been extensively studied. Databases such as the European Bioinformatics Institute and the National Center for Biotechnology Information house these datasets.

Butte: It is funny how many scientists now know how to submit data into these repositories. Very few of them know how to get data out of these systems and know what to do with them. Those scientists now need to get trained in data science and sometimes that means learning a data science language or programing like Python or R. Or maybe it's even simpler. Even dragging these files into Excel is hard for some folks. It is a tough problem. I believe that every graduate student needs to learn a certain amount of data science today.

PNAS: Once a researcher has accumulated some basic computer science knowledge and identified a few tantalizing datasets, he or she will need to wade through the overwhelming amount of downloaded data.

Butte: So you gotta learn some basic statistics. One important part of statistics is how to control for multiple hypothesis testing. You're not just looking at one thing, you are looking at 10s of 1000s of molecules or base pairs or RNA. So you gotta learn those aspects of statistics. And then you end up with a list of findings.

PNAS: Butte says the next bottleneck comes when scientists need to validate their findings by perhaps performing laboratory studies or comparing their findings to other cohorts. From there, scientists may be able to turn the data into clinically meaningful products.

Butte: In my lab we do have a kind of mantra. I believe it that if you want to change the world you can't just keep writing papers about it. But if there's some potential for this becoming a diagnostic or therapeutic, then it's up to you to actually file for some intellectual property there. Not to get rich, but to actually protect it such that it has a hope of getting into a product or service for patients.

PNAS: Butte thinks that big data could change who makes the next biomedical discoveries and where these discoveries are made.

Butte: So it doesn't have to be the big pharma. It doesn't have to be the big funded institutions. I'd like to think, and you know I say this. I…you know…the kid in Bangladesh could come up with the next billion dollar drug if they were enabled by all this data. I think that's possible. At the very least, scientists in training can get to datasets, if you don't have a genome sequencer or any of these kinds of mass spectrometers. You might not have the capital equipment, but there's no reason you can't go get the data today to learn how to analyze that data. And that's empowering I think.

PNAS: Butte wants to make the data even more accessible by creating problem sets or educational material that can be downloaded with the data. These materials would provide a way for professors to add the data into their classes and teach students new research techniques. Thank you for listening. You can find more Science Sessions podcasts at PNAS.org.