Using Data to Fight a Pandemic: A Look at the COVID-19 Research Database

Nacinovich:

For ReachMD, this COVID-19: On the Frontlines, and I'm Mario Nacinovich. On today's program, we'll be taking a look at the recently developed, secure repository of HIPAA compliant, de-identified, and limited patient level data sets that are now made available to public health and policy researchers to extract key insights to help combat the COVID-19 pandemic.

Joining me in this discussion is Dr. Mark Cullen, from Stanford University School of Medicine. Dr. Cullen is the Scientific Committee Chair for the academic arm of the COVID-19 research database. He is also Director of the Center for Population Health Sciences, and holds professorships in the Departments of Medicine, Biomedical Data Science and Health Research and Policy at Stanford. Dr. Cullen, welcome to you.

Dr. Cullen:

Good morning.

Dr. Nacinovich:

To start us off, can you take us back to before the pandemic hit and speak to the already endemic challenges of fragmented data sets and long perspective clinical trials?

Dr. Cullen:

So, I've actually devoted the better part of my last five years professionally to dealing with the issue that you put your finger directly on, which is for those of us in any of the clinical subspecialties or frankly, even in the social sciences, that are interested in the relationship between society, economics, and health, access to data at the level of individuals, so that one can begin to answer the very longstanding, pressing questions about the relationship between how we live and our health, has been extremely difficult. Some of it's been difficult because it's very new and rapidly growing in the area of big data in which we now have access to a wide variety of things, electronic records, and so forth, that were not available previously. But juxtaposed against that, we live in a society which has enormous concerns about individual level privacy. There are enormous concerns about what extents of things might be learned, if the data on people were linked up in such a way that the researchers would have advantage. So, it's kind of a balancing act between getting access to these kinds of data and stepping over a line in which society has been providing reasons uncomfortable.

Nacinovich:

And if we fast-forward back to today, what can you tell us about the COVID-19 research database? What's its mission, and who's organizing this consortium?

Dr. Cullen:

It's being jointly sponsored by a wide array of private sector organizations that make a living in the development and transaction of these kinds of health data. So a group of such organizations, about a dozen, have banded together in an effort to give back, recognizing that they are the holders of information that could be extremely valuable. I'll just mention one particular asset that these companies have, which is otherwise unavailable to researchers like me and my peers, which is that they are seeing the data live, in real time every day, every week, and have the ability to refresh that data. I think the vendors of these various data sets recognized that they were sitting on something that was, from a societal point of view, very precious, which is rapidly refreshed, real time data, that would allow researchers to really look at how the epidemic is spreading, look in very real time at the different trials that are going on, to understand what, if anything, might be working or not working, all of which is probably gonna be relatively easy to observe in these data.

Nacinovich:

Up until now, it's unfortunately been difficult to rapidly answer questions about the epidemiology and treatment of the novel coronavirus. But with the advent of this database, what has changed in terms of data, technology and knowledge sharing?

Dr. Cullen:
Right. So one of the big problems with a data set, once you've stripped off the HIPAA and other legal identifiers – name, Social Security number, address, exact dates and those kinds of things – you could still answer lots and lots of questions, but you can't add more information because you've stripped away all of the details about the individuals that would allow you, for example, to link it to a very different kind of data set, like the data sets many of you have probably seen in recent days on various websites that are following the pandemic. In one way or another, you need to have identifiers somewhere to do that linkage. And yet, those identifiers can never be legally shared with the investigators themselves. So, some of the companies have figured out a fabulous solution to that by encrypting the keys that are the links. No one ever sees the keys except the data managers, who are doing the linkage among the data sets, to make them increasingly useful because they have that characteristic. So this ability, more than virtually any other, is what I think makes the data that have now become available to investigators extraordinarily rich and unique, and eventually even bring their own identifiable data into the trove, hopefully share that data, but in the meantime link their data, if collected, for example, from their own institution, to data that may already be available in this data enclave.

Nacinovich:
Dr. Cullen, what were some challenges that you and your colleagues have experienced along the way in developing the research database?

Dr. Cullen:
We're trying very hard to guarantee that there is no commercial abuse of it. You can well imagine legal groups and political groups and commercial vendors, pharma and so forth want access to this data. So– the companies themselves spent an enormous amount of time thinking out the technical issues, and about those kinds of things. I am now dealing with the next wave of complexity, which is thinking about the actual granular experience that researchers will have. I mean, just literally in the first three days of announcing the website there were 375 people who had registered to get access to the data from literally around the world. But my job is to try and make the friction as low as possible for my colleagues, just to make sure that good ideas get in quickly, that they get access to not only the data, but adequate compute space. These are all things with very real costs, both in terms of the software involved and also the human talent that's gonna be necessary to make it work. So my prediction is that we'll have some lumps and bumps over the next few weeks, as we try and effectively get investigators doing very different kinds of things all onto this platform. So it's, like any other kind of very startup business venture. Of course, we're trying to do it in real time and answer questions that are of great urgency today, and this is not primarily set up for a future understanding. It's set up for trying to change the curve.

Nacinovich:
For those just tuning in, this is COVID-19: On the Frontlines. I'm Mario Nacinovich, and today I'm speaking with Dr. Mark Cullen, from Stanford University School of Medicine, about the COVID-19 research database, which is a collection of HIPAA-compliant patient level data sets, intended to yield insights for combating the COVID-19 pandemic.
So Dr. Cullen, now that we have a better sense of the origins of the database, let's focus on some of the solutions that are now possible. First off, who can access the COVID-19 research database? And how do you anticipate it being used going forward?

Dr. Cullen:
The good news is that we have no specific requirement of any individual who wants to use it, other than that they understand and sign off on the rules of engagement, which have to do with respecting both the privacy needs of these very rich data, but also the promise not to use these primarily for commercial or self-serving or political purposes. But having said that, we would welcome even, for example, scientists from some of our big pharma companies, some of the best scientists in the country work for such organizations. If they're there for any reason other than trying to promote one of their own products, we're all for it. I think the most common users will be coming from the academic sector. So I hope a very broad panoply of different kinds of players. They do, of course, have to have someone on their team that has the technical skills to be able to do the analysis and coding, and so forth, and obviously we hope the database will be used by clinical scientists who are interested in seeing really what does matter, and what does best assure a good outcome when someone does get sick. So we're hoping to be able to answer questions across that whole spectrum.

Nacinovich:
Dr. Cullen, what kind of data and tools are available in the actual database?

Dr. Cullen:
So right now, the first trove of data that will be up and live on the site are what you would call health and health-related data, which is to say there is electronic medical records, so we can see the kind of care and treatment people got for a large swath of the entire

population. There are health insurance claims data, which are now processed much easier to read summaries of every procedure, treatment, and so forth that an individual has had over time. And the advantage of them is you can follow very large numbers of people, and see quickly who's got what diagnosis, when in the scheme of things did they get it in relationship to other things. So it's sort of like a map of someone's health history. There's pharmacy data. So we have access to something in the range, I think, of 90% of all prescriptions that have ever been written in the United States over the last five years. So we can actually look very closely at the relationship between outcome with the disease and outcome from disease, in relationship to drugs that people have been prescribed and are taking. We've got the equivalent of the national death index, so the information on both when people died and what they died of is available to link with the other health data. So that's what's going up first. We're hopeful that we have an increasing amount of transactional data, so that we can actually watch what people are doing all day and have that linked to the health data. We can see what people are spending all day – how much their commercial activity may be related to what's going on in their health so that we can begin to see the interface between the health picture and the social picture.

Nacinovich:
Lastly Dr. Cullen, for any of our listeners out there who are interested in accessing this database where is it available?

Dr. Cullen:
The database is available on the web. I'm always the worst person because I have such terrible online skills here. It's called the COVID-19 Research Database, and if they Google it, I think it'll come up first or second, and they can go on and the first thing they'll be invited to do is to register. And I encourage anyone that's interested in this process, even if they don't imagine immediately potentially wanting to use the data, to register because it provides a window into what we're doing, including questions like, "What are we doing?" And, "How are we protecting people?" In fact, every question you've asked me is on that.

Nacinovich:
Well, that's all the time we have today, but I want to thank my guest, Dr. Cullen, for not only joining me to update us on this massive undertaking behind the COVID-19 Research Database, which is available at covid19researchdatabase.org, but for sharing key insights into the challenges and opportunities experienced along the way. Our support and thanks go out to you and your colleagues, Dr. Cullen. Thank you so much.

Dr. Cullen:
My pleasure. Stay well.

Nacinovich:
For ReachMD, I'm Mario Nacinovich. To access this episode, and others from COVID-19: On the Frontlines, visit reachmd.com/covid-19, where you can be part of the knowledge. Thank you for listening.