

Health Datapalooza Remarks by Dr. Francis Collins, June 3, 2014

Moderator Susannah Fox, Associate Director, Pew Research Center's Internet & American Life Project:

The next speaker is a longtime champion of health data. Dr. Francis Collins is the Executive Director of the National Institutes of Health. Dr. Collins was also one of the leaders of the quintessential health data effort, the Human Genome Project, and is known for landmark discoveries of disease genes. Dr. Collins is the recipient of the Presidential Medal of Freedom and the National Medal of Science, but (I have to confess) that he won my heart when I saw him serenade a meeting of people living with rare disease - Moebius syndrome - whom I consider to be the alpha geeks and the real health care hackers. He sang a song about home and community that brought tears to my eyes, and I tweeted a link to the Youtube this morning, so you can enjoy it yourself. Here's hoping that, although he did not bring his guitar this morning, Dr. Collins will grace us with a few soulful stylings. Please join me in welcoming Dr. Collins.

Dr. Francis Collins:

Well good morning everybody, it's a real privilege to be here at Health Datapalooza; and to tell you something about what's happening with regard to big data - data science - whatever we're calling it today, as regards to biomedical research. I have the privilege of leading the largest supporter of biomedical research in the world - the National Institutes of Health - and I can tell you this topic consumes us, as we gather around the table with the 27 institute directors and myself to try to chart a course forward that will make the most of the exceptional opportunities that now appear in front of us in medical research. And I want to walk you through a few of those areas, in a way that I thought would be particularly relevant for the experts that are gathered here - the 2,000 or so of you - for this Health Datapalooza 5th anniversary.

Well, certainly biological science, which perhaps in the past was considered less of a data challenge, than let's say, cosmology or high energy physics. We have arrived, and boy have we arrived, with a whole variety of types of data. Genomic information, certainly a big part of that, but that's not all of it; other kinds of omic data - imaging - which is often complex, a data requiring much storage space; phenotypes; exposure information; and of course clinical information increasingly derived from electronic medical records. And because the NIH supports research from the most basic to the most clinical, all of these data types have to be a concern for us, in terms of how to collect them, how to store them, and how to make them accessible to the broadest range of biological users.

I'll have to say a word here about the Human Genome Project because it has been, in many ways, a driver of this revolution of data production, and also of open data access. It was my privilege to lead this effort to a happy conclusion in 2003; but it was quite an interesting experience along the way, with many milestones that were laid out and needed to be specifically attained. I just want to draw your attention to one little thing here on this diagram of the first six years of the genome project, with many

milestones as you can see laid out there. And that's this little red box down here in the corner, "Bermuda principles for rapid and open data release established." I think this was a pretty significant moment; although probably not as much in the awareness of many people in this room as it might be. This is a photograph of the people who gathered for that first meeting of human DNA sequencing laboratories in 1996, representing six countries around the world. And here is an actual photograph of what was written on the white board during a discussion we had at that meeting, which was pretty radical. Basically, what we all agreed to is that automatic release of sequence assembly is greater than 1,000 base pairs, preferably daily, and aim to have all sequence freely available and in the public domain. This is something, which had not previously been put forward as part of basic biomedical research. Obviously, the investigators doing this work were basically agreeing, to give away their information without worrying too much about who got the credit because of their sense that this was going to be most valuable as a community resource, that it ought not to be constrained in any way by intellectual property claims or by limitations in access or by the need to pay any subscription fee, in order to be able to see the data. They wanted the best and brightest minds of the world to begin working on it right away.

>>applause from audience

Dr. Francis Collins: Thank you.

>>>applause continues

Dr. Francis Collins:

And that really has started a remarkable movement in biomedical research, which has spilled over into multiple other areas, where the ethic of our enterprise now is – this immediate open access – so that the progress in biomedical research for the benefit of the public can go forward as quickly as possible. And for us at NIH, that has led to a long list of ways in which we have insisted upon that kind of rapid data release and access to the information, with of course, the important issues about patient privacy, which I'll come to in a minute. But many of these datasets are, in fact, such that they do not compromise patient privacy with the appropriate anonymization, and in those circumstances, it is our goal to be sure that immediate access occurs. Well, the rest of the genome project played forward, we ultimately scaled up the effort to get from a very slow, rather manual process of DNA sequencing into something extremely automated, and it was necessary to do so, in order to achieve this goal of reading out three billion letters of the DNA code by 2003. And of course, for me as a physician, and for many of those working on this project, the goal of this is namely, to improve human health, still, was somewhat of a challenge. In fact, various cartoons were drawn, at the time, to document the challenge that we now face, including this one, where you see the caregivers reading out acg, and t, and wondering what the heck do we do with this to try to improve the health of this individual lying in the hospital bed. Of course, having three billion letters of the human dna instruction book, was a milestone, but they're written in a language that we needed to learn how to read and understand, and a whole host of other projects sprung up, immediately; also, placing data in the public domain to try to understand this better. One was the HapMap Project, which basically tried to define where are the common variations in the

human genome, that .1 percent of the DNA, where we differ, and how we could begin to use those to understand risks of disease. Going beyond looking at the variable part, look at the whole thing. Let's not settle for one genome. A thousand genomes project rose up and has now produced complete genomes on about 2500 individuals from around the world. And we needed to understand how the genome actually functions. Only 1.5 percent of that DNA codes for protein. The other 98.5 percent is intensely interesting but difficult to figure out, and it is there where the regulatory signals that turn genes on or off at the appropriate moment are located; and which we needed to understand much better because much of the risks for common diseases lies in that regulatory area. All of those projects involving multiple collaborators and multiple countries, again releasing all of their data, placing it in the public domain as quickly as possible. One thing we have learned, a prodigious amount, in the course of the last five years, is about the hereditary factors in common disease. We knew that diseases like diabetes or cancer tended to run in families, but in most situations, it didn't follow a cleanly understandable inheritance pattern that Gregor Mendel would've approved of. Instead, it was a much messier situation, where there was some kind of familiar risk but you couldn't nail down from previous methods, exactly what was going on.

The genome and the HapMap project made that possible and the strategy actually became a very straightforward one, but one which we had not been able to do before, in a systematic way. And that was, basically, collect DNA samples from affected individuals, collect dna samples from otherwise well-matched unaffected individuals, and then just look through the roughly 10 million places in the human genome, where there are common variants, and see if you can find one like variant B, where there's a skewing of the spelling of that particular variant, in those who are affected versus unaffected. And that tells you, okay, you got a hot spot here. There's something going on with this part of the genome that is a risk factor for that disease. And you can do this systematically and comprehensively. You don't have to guess the answer. That was the big revolution that genomics has made possible, up until you had these complete datasets of information about our DNA instruction book, you had to make a guess, you had to have a hunch. And guess what, most of our hunches were wrong and most of our guesses have been proved incorrect. So to be able to do this systematically, without having to know the answer ahead of time, has opened up many doors. In fact, this strategy, so-called a Genome-Wide Association Study (a GWAS), has now been conducted on virtually every common disease that has a sufficient number of patients to be able to study. And generally, you want thousands and tens of thousands, or hundreds of thousands are even better, and look to see where we are now.

This is a diagram of the human chromosomes (1-22 and the x and the y down there in the lower right). And each one of those colored circles is a place in the genome where there is a variant that's associated - at very high statistical probability - with risk of a common disease. My own lab works on Type 2 Diabetes. There are 82 colored circles there for Type 2 Diabetes; five years ago, we knew about two, and now we know 82. And they point you to places that you never would have guessed that must be involved in the molecular pathogenesis of this disease. Which is giving us entirely new ideas about therapeutics by giving rise, just recently, to an unprecedented collaboration between pharmaceutical companies and the NIH in an open access model to try to use this information to identify the next generation of therapeutics for Type 2 Diabetes. This is a brand new window into what's going on with

enormous consequences but it is a big data challenge to keep track of all of the associated data that you would like to bring to the table, if you're going to decide which one of these colored circles is worth investing a hundred million dollars or more into, in order to try to develop the next drug for a common disease.

Now, of course, how was this data going to be represented? We're here, at a meeting, talking about data science. Of course, all of those colored circles involve DNA analyses on large numbers of people. Yet, we wanted that information to be accessible to those who might have other ideas about how to interpret it because none of these algorithms would be, considered at this point, to be thoroughly mature. Well, here's where a rather novel approach was taken by the National Center for Biotechnology Information. They set up a special database called dbGaP because you wanted qualified investigators to have access to genomic data and phenotype data on individuals 'cause that's where a lot of the opportunity is going to lie. But yet, you wanted to do so in a fashion that respects the informed consent that those individuals gave when they agreed to take part in this research, and many of them had some limitations on what they expected would be the use of the information. So, dbgap is a database where you can see the overall description of what's gone into these various analyses. But if you want to see the data, you basically have to fill out a simple form to say who you are and what you want to do with it so that a quick comparison can be made of your request, with what was, in fact, the consent that the individuals gave, who were a part of that study, to make sure we are not violating our contract with patients, which is to respect their wishes as far as research is concerned. This database has been (I would say) a big success. Basically, those who have downloaded the data have respected the need to keep it confidential. But there's been a huge amount of progress made as a result of the immediate accessibility of the data to those investigators; and yet, we have still managed to maintain the appropriate ethical stance. That kind of situation is going to need to be considered as we go forward in many other ways, where we're collecting patient data; and we need to both advance research and also respect confidentiality and the consent process that was used. Now, this is all about, actually, measuring that .1 percentage of the genome, where you see common variations; but we're getting much better at sequencing the whole thing. Look to see here what's happened, as far as the cost of obtaining a complete DNA sequence on you or me; which, back in 2001, would've cost you about 100 million dollars and now, as you can see from this curve, a dropping very profoundly, moving from sequencing machines that looked like a phone booth to some of them, that now look like postage stamps, like the one I'm holding up in my hand. This is the Ion Torrent, which can sequence a genome in two or three days and we haven't reached the end of this. Right now, we're down to about \$4,000. There's been a projection that by the end of this year, one of the instruments now available will drop the cost of a complete DNA sequence to \$1,000; which has been sort of a mythical goal to reach out there. That really does put us in a situation, where increasingly, it will be attractive and appealing to obtain that DNA sequence accurately and confidentially, once and for all, placing it in the electronic medical record, where it can be immediately available when a decision needs to be made. As for instance, a doctor is trying to prescribe a drug and wants to know whether that drug is the right drug at the right dose for you. And there are more than 100 such drugs, now, that have on the FDA label, an indication that physicians would be well advised to check the patient's genotype before making the prescription. And yet, very little of that is happening because of lack of immediate accessibility to the

information when you want it. If the information is already in the medical record, it's a click of a mouse to do that - and this whole field of pharmacogenomics which has been a bit slow to find its way into the main stream of the practice of medicine - we'll really have a chance to do so.

Of course there's a lot of data here that's being generated; and, if I had more time, I'd talk to you about where things are going with cancer because cancer is a disease of DNA. And it is increasingly the case that if you want to understand cancer in an individual, you don't really want to depend on what those cells look like under the microscope - whether they're big or small, or blue or red, or whatever. What you really want to know is what genes are mutated in that cancer and how would that play out, in terms of predicting response to therapy. And yet, all of the data that's being generated in cancer genomics - from many different sources, needs to be assembled in a place (again), where all the bright minds of the planet can compute on it, in order to understand some of the nuances that are still waiting to be discovered. In that regard, you might want to know about a new enterprise, called the Global Alliance for Genomics & Health (GA4GH), which has as its mission, to accelerate progress in by helping to establish this common framework of harmonized approaches. This is not, itself, going to run a database or put data into the cloud; it is establishing the kinds of standards that we want to see attached to any such effort so that data sharing can be achieved at maximum speed and effectiveness. There are already 183 partner organizations and 26 countries represented. You can read more about this at the GA4GH website. NIH is strongly supportive of this but we are not driving it; it is very much bubbling up from the scientific community and the advocacy community, as a need that needs to be met.

Now the data that, of course, is being generated is not just from genomics. If you are watching what's happening, in terms of the uses of big data, in a place like the National Center for Biotechnology Information at NIH, it's pretty breath-taking. This diagram shows you the number of daily users at NCBI, over the course of the last 14 or 15 years; and you can see we're up now to 4 million daily users, daily downloads of 35 terabytes. We have certainly arrived here in the big data zone. NCBI runs the genomic databases that many people use; although, they are displayed in other places and it's good to have some good competition here. NCBI also runs, of course, PubMed. PubMed, the place where NIH investigators are required to deposit their publications in PubMed Central after twelve months, and increasingly to deposit them as soon as possible, in fact, immediately, on publication. We are big proponents of the importance of having access to scientific publications that the public has already paid for and shouldn't have to pay again in order to be able to see that information. And, I think we're making real progress here; although, we're not completely out of the woods. There are still moments where efforts are made to try to block that kind of open access for other purposes that we think are less noble. And again, I would encourage you - if you're engaged in this - to continue to make the case for why it is that scientific publications ought to be made publicly available, without other kinds of blockades. I think we're not completely clear, at this point, that this is a solved problem.

So, I only have a few more minutes left. I want to tell you a little bit about some of the things we're, specifically, doing right now to try to meet this big data challenge. I assembled an advisory group, two years ago, as a working group of my advisory committee to the director; and, I asked them to review NIH's current plans for managing the big data challenge and to come back with recommendations and they did. They came back with some very strong recommendations. One of them said, if you don't do

something quickly to address this issue, you will be committing institutional malpractice. Those were strong words (that got my attention) but believe me, we are taking this with the greatest seriousness. So here are four things we're doing:

- I established a new leadership position – an associated director for data science, who reports directly to me, and I think for many other scientific agencies, this is something we need to do in order to address the urgency and the importance of this issue, at the highest level. And I was successful in recruiting a remarkable data scientist from UCSD – Dr. Philip Bourne.
- I established a scientific data council – made up of representatives from all of the 27 institutes and centers – and an external advisory board with experts from the outside.
- We started a brand new effort called Big Data to Knowledge - which is a research-oriented enterprise - to focus on data sharing, where you try to get the data that is often-times is hidden away somewhere into a place where it can be found using a data discovery index, and to establish better standards.
- We are investing, more heavily than ever, in software development and in training; and we will be funding, in the next couple of months, the first Centers of Excellence in Data Science. And we got some very exciting applications, which have been very well-reviewed, and it's going to be tough to decide in the current budget climate, which of these we can afford to start funding. But, we will make the best choices we can.
- The Data Commons is an area, which Phil Bourne has been very much investigating and proposing; but yet is not completely formulated, in terms of what it must look like, other than the principle. The principle is that we ought to have a place – a virtual place – where data that people are looking for can be found; supporting sharing, accessibility, discoverability; enabling scientific innovation by having this co-located with advanced computing resources; putting all of this in a cloud-based computing environment. Again over the course of the next few months, there will be much more details emerging about what we want to do with this data commons concept, but I hope you would resonate with the idea of the importance of making this in a fashion that is no balkanized, but is accessible in a place that's readily usable and associated with appropriate standards.

Two other things I want to say before I stop. One is a focus that we have right now on ClinicalTrials.gov. This is the website where information about current clinical trials is posted across the world, and it is in fact, a very useful place for patients to go and find out what studies are underway. But it is also now, as required by the FDAAA act, a place where we need to have results posted and we are in the process of documenting exactly what that needs to look like, in order to enhance access to everyone about what has happened as a result of a clinical trial – whether it's cancer, diabetes, heart disease or whatever; so something that might be well-worth watching. Finally, one other thing that I don't know whether it has been mentioned at this meeting because I was not able to be here yesterday but I want to put it in front of you. As a new development in the United States, which I believe is powerful and unprecedented in its ability to generate information about what works and what does not work, in the real world of medicine. (Sort of coming back to what the preceding speaker pointed out as an area, where we desperately need more data). And this is the formation by the Patient-Centered Outcomes Research

Institute - on whose board I sit and which was formed by the Affordable Care Act - of a National Patient-Centered Clinical Research Network, to be called PCORENET. This is constituted of 11 healthcare delivery systems and 18 patient-powered research networks, collectively representing 26 million patients across the country. Allowing one, once this is fully set up with interoperable databases and electronic health records, to be able to conduct observational trials almost immediately and almost for free and also providing a platform for randomized-controlled trials of unprecedented size, scope, and potential, reasonable cost characteristics as well. This, I believe, will be a major advance for our country and the ability to conduct these kinds of research studies in the real world and I think everybody should be watching closely as a place where we can test out lots of things, including mobile health applications, by the way.

Well, this meeting, very much, talking about data science, I think is trying to make the case that we can't necessarily predict exactly where this is going, nor should we try. And I'm sort of fond, therefore, of Antoine de Saint Exupery's recommendation here: "As for the future, your task is not to foresee it, but to enable it." I think that's what Health Datapalooza is all about; that's what we at NIH are trying to participate and assist with, with the remarkable talents of our investigators across the country and across the world. And, it's been a great privilege to be part of this exceptional meeting, here, this morning.

Thank you all very much.

>>>applause from crowd

Susannah Fox:

Thank you Dr. Collins; and I would just like to say that if there's anybody who's enable thing future, it's Dr. Collins.