Mr. Spradlin, the floor is yours.

Dwayne Spradlin: Thank you. Hello, everyone, and welcome to today's webinar entitled "Big Data -- Exploring New Models for Research that can Improve Health and Health Care." We are thrilled to have you join us for this session, which we hope will be both informative and thought-provoking. With a registration of nearly 600, this is clearly an area of intense interest.

My name is Dwayne Spradlin. I'm the CEO of the Health Data Consortium, a public-private partnership based in Washington, D.C. dedicated to open health data and mobilizing health data to improve health and health care in the United States in areas from improving clinical care to enhancing medical research to empowering consumers with knowledge and choice. I will be your moderator for the next hour.

This webinar is part of the Innovators in Health Data Series, on which we partnered with PriceWaterhouseCoopers. Health Data Consortium, or HDC, and PWC, our goal with this series is to bring top leaders and thinkers to you to share their work and inspire better and broader use of health data and all the benefits that will come from that. And, of course, this particular webinar would not be possible without the support of the NIH and PCORI, for which we are thankful.

We are obviously thrilled to have three special guests join us today. Patient-Centered Outcomes Research Institute, or PCORI, Executive Director, Dr. Joe Selby; National Institutes of Health Director and PCORI Board of Governors member, Dr. Francis Collins; and NIH Associate Director for Data Science, Dr. Philip Bourne.

Each of these guests are renowned and leading figures in this space. Drs. Selby, Collins, and Bourne will each present, after which we will address a number of questions to the panelists including some from you, the audience.

You will notice the Q&A box on your screen. Please feel free to enter questions at any time. You may also tweet questions to our Twitter handle, @HBConsortium. While we will not be able to address all the questions, we will do our best to address as many as possible.
Finally, please note that this session is being recorded and will be available later for replay.

With that, it is my distinct honor to introduce our first speaker, Dr. Joe Selby, a family physician, clinical epidemiologist and health services researcher. Dr. Selby has more than 35 years of experience in patient care, research, and administration. Since July 2011, he has been the Executive Director of PCORI, which was established with the Affordable Care Act to fund Comparative Clinical Effectiveness Research. Today Dr. Selby will be telling us about how PCORI is looking to leverage data in its work including its development PCORnet, the National Patient-Centered Clinical Research Network. Dr. Selby, the floor is all yours.

Dr. Joe Selby: Thanks, Dwayne. Thanks very much to you and Health Data Consortium for the invitation to join this interesting conversation. It's a privilege to be on the same panel with Dr. Collins and Bourne.

So -- PCORI is the latest, I would say, in the array of funders who have made major investments in fleshing out a vision that is just irresistible, and that is the vision that big data, whether it comes from previously conducted research or whether it's generated in the process of delivering health care, can be harnessed to answer questions of critical importance for patients who currently don't have satisfactory options or currently have multiple options and are having trouble choosing which one works best for them.

PCORnet is $100 million investment that PCORI made at the direction of its Board of Governors in December of 2013, and I am going to tell you a bit about it today. Distinguishing features of PCORnet that we see for the first time -- there have been previous large big data health care delivery-based networks, but PCORnet is the first one that starts from the electronic health record. Now that it has been disseminated widely across the country, that's feasible.

Most previous networks were based more on what we call claims data, the data of the transactions and actually often the bill-paying associated with health care. We are starting in a different place. Those of you who know Health Informatics know that working with electronic health record data is challenging. I'll say a bit more about that.

And the second distinguishing feature is PCORI, in part, because of its name and, in part, because we've become convinced that it's the only reasonable way to do research. It spends a lot of time and effort engaging patients as well as clinicians as well as the delivery systems that generate or oversee the collection of these data in not only the governance of the network but its uses. And, in our view, that's the way we ensure that meaningful research gets done and that it gets used.

Indeed, we see here PCORI's mission statement, this was crafted back in 2012 by our Board of Governors, and it simply says that we do research, the bolded parts is what I want to draw your attention to -- "Research that is guided by patients, caregivers, and the broader health care community." So it is in our name, and it's in our DNA that we put patients and stakeholders at the center of the research activity. And PCORnet is not an exception. Next slide.

PCORI's Board of Governors also laid out three strategic goals. These are our only strategic goals, but they are pretty challenging. The first one is to increase the quantity
and quality and timeliness of comparative clinical effectiveness research information. That's our mandate in the legislation.

The second goal is to speed the implementation and use of that evidence. We don't want it sitting on the shelf. We want it changing practice and improving patient outcomes. That's also in our legislation.

And the third is to influence research funded by others. So in our view, all research or all clinical research ought to be patient-centered. It ought to use the best methods for comparative effectiveness research. And so those are our three goals, and my point here is that PCORnet really addresses all three of those. Next slide please.

PCORnet is a network of networks, if you will, and I'll go into that a bit more in just a minute. But it has two goals. The first is to improve the nation's capacity to conduct clinical research more efficiently by creating this large representative research network. And ours is focused on comparative effectiveness research, although I hasten to add that having built the network, it could be used for a variety of purposes in addition to comparative effectiveness research -- safety research, even idiologic research, and, certainly, systems improvement research.

We see our mission, and this also is a distinguishing feature from many previous networks as critically important to support randomized studies as well as observational studies. So previous networks often focus just on using the data to get relatively quick answers and observational analyses. We believe that much compared to effectiveness research does and will continue to require randomization. So PCORnet emphasizes the capacity to do randomized studies.

At the same time, we want a network that supports a learning U.S. health care system. And what we mean by learning health care system is turning out to be a pretty darn interesting question. In some people's view, a learning health system is simply a system that keeps improving itself and keeps making it possible to deliver higher quality care more efficiently.

To others, a learning health system means as we deliver care, we learn more about the causes of disease and the treatments for these disorders. And that's a tension, actually, that PCORnet will be working with as we move forward. But we definitely think that the answer is we can have both in one network. Next slide, please.

So this is the main centerpiece of my presentation. This is a picture of PCORnet, and in the middle you will see a blue circle signifying the members of PCORnet. They are 11 clinical data research networks. These are systems, each of which must accumulate complete clinical data on at least 1 million patients. These systems range from systems like Kaiser Permanente and OCEAN, which is a national network of federally qualified health systems, to the city of New York, the state of Louisiana, the city of Chicago, and several networks of academic medical centers who are banded together to create "systems."

So those generate -- those start from electronic health records, and aspire to generate complete health data. Sprinkled in among them are 18 patient-powered research networks. These are groups of patients with single conditions who have banded together on their own or with the collaboration of a researcher or an advocacy organization because they are interested in pursuing research. And we actually see these patient-
powered networks as maybe, in some sense, the spokespersons for patients. In other
senses, the innovators and maybe even to a bit of an extent, the troublemakers.

So this is a part of what's going to keep PCORnet actually lively and jumping as we go
forward in time representing the needs of patients to harness this data and financially to
the questions they've got.

I want to point, for a minute, to the light blue box on the left. That is our steering
committee. We've taken care to put on the steering committee in addition to all the
network representatives, representatives from other entities that will, we hope, someday
use PCORnet as a site for research. And we start with the NIH, the FDA, ARC, and the
CDC.

You will also see that CMS, the Office of the National Coordinator and the Office of the
Assistant Secretary for Planning and Evaluation are on the steering committee because
they hold the keys to accessing CMS data, which is going to be crucial to really complete
the data for many of the patients in both of these types of networks.

Also on the steering committee are representatives from medical products; that is, Arma
and device manufacturers, because we fully intend, we hope, that PCORnet will be seen
as an attractive place for those funders to cite some of their research, and to find an
efficient home for their research.

Now is the real work of PCORnet. PCORnet, I should say, is funded that $100 million-
plus will be invested over the next 18 months, we're about three months into it, and these
are the tasks that it must accomplish. I'm going to just go through them briefly.
Governance means how do we make decisions about what research will be conducted and
what research will not be conducted. How does an individual network decide which
projects it wants to participate in and which ones it's not interested in participating in.
What are the rules that govern how much research you must participate in to be part of
PCORnet and to receive the PCORnet benefits? So that's one task force. Each one of
these, it will strike you, is an immense amount of work.

The second is data privacy. It's only second on the table, but it's first, really, in our
minds. If we don't ensure the community, the patient community, the health care delivery
system community, that we are putting patient privacy and data security first, we'll fail.
This is a fascinating dialog with patients, really, on both sides of the fence that the data
should be made more open and the data should be held more tightly, and we believe that
we can do both.

The third is ethics and regulatory, and this speaks to the critical challenge of bringing the
Institutional Review Board to the human subjects research committees into the 21st
century, into a state of awareness about this type of research and the differences between
this kind of real-world low-risk research and the kind of research that we think of when
we're trying to understand the causes of disease or early stages in the development of new
therapies. It's not that we will never do those kinds of research within PCORnet but there
are distinctions, and IRBs need to find ways to streamline the regulatory oversight if this
research is ever to be really effectively and efficiently conducted.

Data standardization -- no two data systems within PCORnet are exactly alike. They
generate data using different expressions or different standards. But they also generate it
with different motivations. So you can have data that look the same but are not. So we
have to not only standardize the data but then scrutinize it closely to make sure that it's really comparable.

Health system interactions refers to the fact that we need to engage the health systems that are generating the data very actively. We need to make them appreciate the values of participating in a network both because of the insights that can be gathered for running their businesses, but also for the collateral effects of improving the quality of their data and learning from more basic or more upstream research. That is a very intriguing challenge.

Engaging patients and consumers, and I want to add engaging the clinicians in these systems is another area where we have a task force dedicated.

Preparing to conduct clinical trials is a huge step. Many of these systems have not hosted clinical trials, or they have hosted very small and inefficient clinical trials, and we're trying to help them gear up to be much larger contributors to clinical trials -- clinical trials that address questions they are interested in.

We have an unprecedented opportunity to address rare diseases just because of our size. We'll be about 26 million lives when the 18 months are done, so that creates opportunities to study rare diseases. Many of our patient-powered research networks actually reflect patient communities of those with rare disorders.

A number of our networks already have bio repositories, and that is a goal for all of them. These systems generate specimens, and we need to capture them and the permissions from patients to use them.

And the last one is an obesity cohort, which is really a trial balloon, a proof-of-concept. We're asking every network to create an obesity cohort.

I want to emphasize that PCORnet is not a single large database. We don't transmit all the data to a central site. The data stay within each individual network, and are accessed by centrally written programs. That emphasizes the importance of having a common data model in the first place. Next slide.

This just illustrates the coverage. The blue triangles are the CDRNs, and you'll see that they gather data from most states in the union, and then sprinkled in alongside them are the green circles, which are the PPRNs, and they, again, often, are multistate in nature. But we really have -- certainly don't claim to have a country covered quantitatively, but in a representational way we are geographically very nicely dispersed. And then the last slide, please.

So where are we going to go -- where does PCORnet aim to arrive by the end of 18 months? Well, each of the clinical data research networks needs to have a defined set of standardized clinical data that's interoperable with data from the other CDRNs, and each PPRN, similarly, has to have a standard database, and these are probably going to have varying amounts of clinical and patient-generated data depending on how much progress they are able to make in getting at the clinical data. I think this will vary across the PPRNs.

But what data they have will be stored in the same common data model architecture as that of the CDRN, so that they can interact. PCORnet will have clear policies on
decision-making, on uses of the data, on how we collaborate and share knowledge across the networks and also on -- and with the external researchers as well and through policies on data privacy and security abetted by the patient community.

Within each participating clinical data research network, we will have a clear representation of patients, clinicians, and health systems themselves not only in governing the network but in using it. The CDRNs and the PPRNs will have progressed to a point where they can participate in both large observation studies but also, and importantly, in large, typically, rather simple randomized clinical trials. And the networks will demonstrate a readiness to collaborate with outside researchers. That's part of the proposition that (inaudible) resources and means this network is open to researchers outside of the network as well.

And, Dwayne -- let's see -- I believe that's my last slide. Dwayne, I'll stop and I look forward to Dr. Collins and Bourne's comments.

Dwayne Spradlin: That's terrific, Dr. Selby. Thank you very much. That was both extremely interesting and visionary. So now our next panelist will be -- that I'll present -- he is Dr. Francis Collins. Dr. Collins is the Director of the National Institutes of Health. In that role, he oversees the work of the largest supporter of biomedical research in the world, spanning the spectrum from basic to clinical to population health research.

Dr. Collins is a physician geneticist noted for his landmark discoveries of disease genes and his leadership of the International Human Genome Project, which culminated in April 2003 with the completion of a finished sequence of the Human DNA Instruction Book. Today Dr. Collins will be discussing the different types of data, growing biological databases in the NIH's big data challenge. Dr. Collins?

Dr. Francis Collins: Thanks, Dwayne. It's a pleasure to be on this webinar with my friends, Joe Selby and Phil Bourne. I thought I would follow up with what Joe nicely presented about PCORnet, which is an enterprise that I am very excited about and privileged to serve on the PCORI Board with some general reflections about the ways in which big data has become a very prominent part of the agenda for anyone interested in biomedical research.

We've clearly arrived at the point where the amount of data that's possible to generate is growing with extreme rapidity, and the challenge to all of us is to figure out how to make the most of that, make it accessible, make it computable, put it into formats that people can actually be able to make sense out of.

This is one of the reasons that I recruited Phil Bourne, who has been here at NIH for all of a couple of months as my Associate Director for Data Science, and you'll hear from him shortly about his perspective about some of the big picture items that we're talking about here at NIH. So if I could get the next slide.

I think it's probably not really necessary to go through all the points about the value of data sharing because most everybody now accepts them and people on this webinar, no doubt, are pretty much bought into them. And I will tell you, certainly, from my experience as a leader of the Human Genome Project, one of the most critical decisions we made back, now, almost 20 years ago was the decision to make all of the DNA sequence data from the Human Genome Project immediately accessible on the Internet with no barriers to access so that everybody with a good idea could start using it.
And that really was, I think, a signal moment and has led down through several iterations after that to a general conclusion that if we were talking about data sets that have enormous potential for value for many users that the goal ought to be to maximize transparency. It helps to validate findings, it facilitates additional research, and increases the return on investment.

So at NIH we are engaged in many ongoing efforts to increase and facilitate data sharing. You'll hear about BD2K briefly from Phil and, certainly, I want to say some things about our plans for increasing public access to data. Next.

Again, the amount of data that's getting generated can be depicted in various ways. This graph shows the daily use of the bio repository data that NCBI, the National Center for Biotechnology Information, contains. And, of course, this includes such things as genome sequences and PubMed daily page views -- 28 million daily downloads, 35 terabytes. And as you can see, the users going up much more steeply than a linear course.

So, yes, we have arrived in the same zone as the cosmologists, the high-energy physicists, the weather predictors, and a few others for whom terabytes have been a reality. We are there now, and we're headed, maybe even faster than the rest of those fields in terms of the accelerating pace of data collection when you consider genomics, electronic health records, imaging, and so on. Next.

I thought I'd give you three areas of data sharing, both challenges and solutions. This is not intended to be an exhaustive list of the things under consideration at NIH. Many of these intersect with what you just heard about from Joe in terms of PCORnet. I thought I would touch upon them because they're timely. So next.

The genomic data sharing, next, is a reflection of the fact that the amount of genomic data now possible to produce has been escalating so that it is possible for single investigators to actually generate a huge amount of DNA sequence information. This chart shows you how the cost of sequencing a humane genome has dropped from about $100 million in 2001 now down to about $4,000 and aiming to get through the mythical $1,000 genome, perhaps even by the end of this year.

That means that sequencing centers, whether they're working on cancer or rare diseases, common diseases, are capable of generating huge amounts of this information and, obviously, it is of greatest use if many users can have access to it.

The next slide shows you a summary of where we are right now with an NIH genomic data sharing policy, which we intend to implement in its final form by January of next year. Again, for our grantees, it expands the expectations that they will share genomic data under the policy that was developed for genome-wide association studies, GWAS, but it takes that policy and extends it also to large-scale non-human and human genomic data not just snip chips.

This is intended to ensure broad responsible sharing. Investigators who submit data are now needing to provide their data-sharing plan to us when they apply for a grant, submit their data in a timely manner, living up to that data-sharing plan. But, obviously, we want to be sure to be consistent with consent. So for human data -- and I'll come back to this in a minute about the Common Rule -- to obtain consent for data that will be used for
future research purposes because otherwise the limits put forward by the lack of consent can really get in the way of the value of the enterprise.

Responsibility of investigators who accept and use the data is to follow terms and conditions of the research use, which means also recognizing that those who have produced the data should have the chance to first publication on it.

So this is something to watch for. It is fairly mature at this point. It will certainly have an effect on the accessibility of genome data, at least coming from our grantees. Next slide.

There is certainly going to be more, not less, of this given that FDA has now cleared two tests that use this illumina MiSeq diagnostic platform. These are two tests for cystic fibrosis -- paves the way for more genomic technologies. So you would expect, therefore, as part of clinical trials and other things, that the amount of DNA sequence data that finds its way into accessible databases should go up.

Now, this might cause you to worry a bit about how is all of this going to be collected in some sort of standardized format and placed in some sort of accessible location where it can be queried by interested users.

The next slide is just the logo and the name of a new organization, the Global Alliance for Genomics and Health, which has been organized primarily by academic researchers who are concerned about this issue, particularly as it regards cancer genomic data. But this is now international and now involves participation by funding agencies and by industry; the idea being to set standards for how genomic data should be, basically, formatted and deposited in the cloud in such a fashion that interested investigators have maximum abilities to be able to use it effectively. And this is certainly an organization to watch. If you go to the Web, you can read more about GA4GH, as it is called. Next slide.

The second theme I wanted to mention is clinical data sharing -- what about that part? We are, of course, anxious to see that as much as possible of that information finds its way into public view. And here I have to say we have some concerns about the way in which NIH-funded trials in some instances don't find their way in the public view. This is a graph showing you, on the basis of an analysis published in the British Medical Journal, about what is the percentage of studies that get published over the course of time from study completion. You can see on the X axis there going from zero up to 100 months.

And even after 100 months, you can see that about 65 percent or so of these studies have published, and some of them still have not and probably never will. So we have both a delay problem, and we have an incompleteness problem, and this is something that we want to take seriously.

The next slide shows you an analysis that was done by investigators at the Heart, Lung and Blood Institute and published in the New England Journal. Again, looking at publication rate of clinical trials and noting, as you can see here, that for those trials that were basically designed to measure surrogate endpoints, the results are not as impressive in terms of cumulative publication rate or the timing of that, whereas, if clinical endpoints were part of the trial, things look somewhat better, and there's certainly a lesson there as well. Next slide.
Basically, the way in which a lot of this problem is going to get addressed is by clinicaltrials.gov, which is the website, which you can go and look at right now if you're not familiar with it, which lists all ongoing clinical trials. But now, as a consequence of legislation, will also require those running the trials after completion of the trial to be able to deposit in there information about the data that has been generated.

This is an intention, therefore, to increase transparency, and this is, in fact, something we're in the middle of figuring out exactly how to implement with the legislation that requires this, which is in the FDA legislation generally known as FADAH (ph). So also this should be a window into the details of clinical trial data. In this case, not based on individual patient data but on aggregated data but still in very useful format.

A final thread that I wanted to play out for you, next slide, is Human Subjects Protection, which, of course, interdigitates with the first two threads I've been talking about. Next slide.

We are in the midst, I think, of trying to put forward the first really significant revision in the Common Rule, which oversees how human subject research is carried out since 1991, as there, obviously, have been huge changes in research since then, growth and research volume, increase in studies that use multiple sites, increase in health services and social science research, new technologies, the private sector involved in collaboration sharing of specimens and data. All of these putting stress on the current system, which was not well designed for this situation. So -- next slide.

Outlines what some of the proposed reforms are, and these were outlined in opportunity for public comment back in July of 2011 in something called an ANPRM, an Advanced Notice of Proposed Rulemaking. Part of this is enhancing protection, and that would include, going forward, requiring consent for research with biospecimens, recognizing that DNA sequence data is unique to the individual. And it's increasingly difficult to argue, prospectively, that such samples can be anonymized or de-identified; enhancing data security and information protection standards, another part of that; and to be sure that this is extended to all research conducted at federally funded institutions not just federally funded research.

But there is also a need to reduce the burden because the amount of paperwork and bureaucracy that have crept into oversight of human subjects without any real benefit as far as most people can tell, to either the patients or the researchers, that requires a reconsideration of the implementation of the Common Rule promoting a broader consent of future research with biospecimens and data, broadening the exemptions for low-risk research, which currently are subjected to the same very stringent oversight as something that is potentially really risky, and doing something about the remarkable redundancy when you have a multi-site study of multiple IRBs basically requiring that such studies operate from a single, central IRB.

That has been the subject of much public discussion. The next iteration of this, the so-called NPRM, should be appearing sometime later this year with hopes that this will be, in fact, in place by 2015, in time to facilitate a lot of the research that's going on for large-scale and small-scale human subjects efforts. And, certainly, this intersects in ways that are pretty obvious with what Joe was talking about with PCORnet.
Again, this is a bit of a spotty representation of some of the things that are on our agenda here at NIH about big data, but I'm glad Phil Bourne is going to speak next and can talk to you a bit more about some of the things we're trying to organize as far as a trans-NIH focus on big data and how to make the most of this digital enterprise.

I'll stop there and turn it back to Dwayne.

Dwayne Spradlin: Thank you very much, Dr. Collins. Let me now introduce Dr. Philip Bourne, also of the NIH. The first permanent Associate Director for Data Science, Dr. Bourne is leading the NIH-wide priority to take better advantage of the exponential growth of biomedical and population health research data, which is an area of critical importance to health-related research. One area to which he is extremely committed is the furthering of free dissemination of science through new models of publishing and better integration and subsequent dissemination of data and results. Dr. Bourne, take it away.

Dr. Philip Bourne: Thank you, Dwayne, and thank you, Joe and Francis, those excellent talks. I'm just going to, perhaps, drill down a little on some of the things that Francis was describing to really give you a sense of where we're going at the NIH. And as Francis also pointed out, I've only been here a few weeks, so these are certainly early developments that, I must say, I'm having a great deal of fun with it. Can I have the next slide, please?

I'd like to start with some of the observations that I've made, and both, I would say, good news and bad news. The good news is why we're all here, in fact. I think the notion of better and more data sharing offers unprecedented opportunities, and I think we're all here to try and figure out the best way to make that happen.

We're beginning to plan for that and quantifying what needs to be done. Of course, this is not being done in isolation. This is being done by using an amazing group of scientists that we have in our (inaudible) community and also, of course, within our intramural community notably at the National Library of Medicine and CBI.

But it's also, I'd say, some of the observations have bad news. It's clear that if you look at data growth that's being described here today against the budgets that we have, this undoubtedly represents a significant sustainability challenge, and we have to come up with ways of addressing that.

At the same time, from various directors that from the Office of Science and Technology Policy, for example, what's been handed down is essentially the why for doing data sharing but not really the how. And that is, obviously, a major consideration to people like myself.

But, having said that, we also, in a situation where we really, you know, we're talking about big data going forward and so forth, but we don't have the best possible handle on how we're using the data that we currently have, and I think we're in the process of beginning to quantify that and understand more about current usage patterns, which will help inform us in what we need to do, going forward. But, having said that, of course, it's difficult to estimate what the supply and demand is. Next slide, please.

So in thinking about these kinds of issues, we've come up with five programmatic themes and the associated deliverables, and I'm going to have -- next slide, please -- and I don't have time to tell you -- next -- about them all, but the vertical columns through these slides really talk to the themes. One is, of course, is sustainability.
Another very important one is education. It's clear, wherever I go, people inform me that we must do more in training by medical data scientists and, certainly, we have plans afoot for that and training grants that are already going out there.

Innovation is another area, and then what processes we use to handling data-related activities internally, and then another important one is collaboration, both nationally and internationally between our various federal partners and partners overseas. Data, of course, is very much an international activity, but it tends to be funded nationally, which creates some interesting challenges. Next, please.

I'm just going to tell you very briefly about two of these activities. One -- and deliverables associated with them -- one is what we're calling The Commons, and the other one is, as Francis mentioned, the Big Data to Knowledge Initiative. Next, please.

So -- next, thank you -- so just quickly going through this is we have data, we have the why that's been defined. We need to get to the how and then, of course, is the end game. And as you go on, the data comes from a variety of different resources -- sources, I should say. Obviously, the clinical patient, which is a major consideration here, but there's also data that comes -- a large amount of data that comes from core facilities. And then there's the long tail of science -- a very large number of researchers that produced by, in relative terms, small amounts of data. But when it's put together, it represents a very large amount of data.

So how are we going to begin to address this in a way that's collaborative and also provides an opportunity for interoperability and interaction, and so on. And next -- and we have various stakeholders -- keep going through these, thank you -- we have the NIH awardees, government, academia and the private sector. And what we're putting together is a series of initiatives around BD2K, which I'll get to in a second, but these all fit together in what we're referring to as The Commons, which is really centrally a place where this can be gathered, various research objects and data, software, and so forth, can be indexed and used in accordance with business models that are sustainable. Next, thank you.

So what will The Commons accomplish? Community building through sharing and accessibility and new models of innovation for how we manage data and achieving this balance between accessibility and privacy; cost effectiveness, which is, essentially, a significant problem right now, in my view in terms of the cost of multiple sites effectively doing very similar things; and it provides opportunity for collaboration across boundaries. Next, thank you.

So The Commons will be empowered by the Big Data to Knowledge Initiative, and there are a series of calls and opportunities for data scientists to participate in this. So, for example, data has to be discovered before it can be used, so there's a callout, which is currently under review, for a consortium to develop a data discovery index, which will enable the discovery and use of data.

There's programs for establishing and maintaining metadata standards; software maintaining and finding software; training centers and grants with a goal of, really, providing more trained personnel but also to improve the recognition and reward system that goes with being a data scientist in the biomedical arena. And then, also, lastly,
censors (ph) that will engage a large fraction of data scientists throughout the community to work on these problems.

So this is just the beginning, and you'll see a series of calls and opportunities in the months to come where we'll begin to actually make these things happen.

So I'll stop there, and I'll hand it back to you, Dwayne, and thank you.

Dwayne Spradlin: Thank you, Dr. Bourne, and to all of you for those excellent presentations. This is obviously exciting, exciting work. Of course, you all know the Q&A can be as informative as the presentations themselves, so let's move on to the Q&A section of the agenda. And remember, listeners, that you can ask your questions in your Q&A box at any time or send them to our Twitter handle @HDConsortium. So let's go ahead and get started.

So for my first question, let me direct that to Dr. Selby. Dr. Selby, why does the mission need PCORnet now, and why isn't something like this in place already?

Dr. Joe Selby: Well, I think these three presentations have kind of coalesced to indicate the potential that becomes a need. I mean, once you see it, it essentially is a need. We need it because research costs too much, therefore, most of what we do we do without good evidence. More than half of our clinical decisions are made without evidence. So we need it. We need it because we are practicing a lot without this evidence and just studying the consequences of what we do could advance us along the road toward, number one, better outcomes, better consequences for the choices we make.

But I think there is really, also, a case to be made that if we are, on the average, making better choices, we will save health care resources as well. Certainly, we'll save resources by not wasting care, and we'll save resources by not doing care that harms.

And why now? I think the emergence of the electronic health record actually brings this - makes this even more urgent. I mean, the amount of information that's now digitally stored is essentially limitless and will be a long time in cataloging and mining that. It's clearly time to get started.

Dwayne Spradlin: Great, thank you. Okay, the next question I'm going to direct to Dr. Collins. Dr. Collins, you talked quite a bit about the emergence of vast sums of data. Some say that more data has been generated in the last two years than in all of prior human existence, and a lot of that is health-related data. Could you talk for a moment about what kinds of problems we think potentially advance or solve with these huge volumes of data that we couldn't otherwise?

Dr. Francis Collins: Well, I think what we're really interested in is trying to identify what are the causes of illness and what are the effective preventions and interventions that successfully can prevent or treat those diseases. And, oftentimes, you need very large data sets of multiple types to do that. If you look at the medical literature now, I just sort of noticed -- I wrote a blog about an example -- Atul Butte, who is an investigator at Stanford -- by looking at publicly accessible data about genetic factors that play a role in disease or various serum chemistry markers, came up with a hypothesis that there might be a connection between some of these things and then was able to test it using electronic medical record information from three academic centers that he had access to.
And came up with five observations published in this paper in Science Translational Medicine, about potential biomarkers for disease. For instance, magnesium is a biomarker for gastric cancer. Who would have thought it? It's that kind of ability to take very large quantities of data, not necessarily to start out with a narrow hypothesis but to see what the data can tell you about correlations between various features that puts us at the threshold of new insights into medicine.

Obviously, you are more empowered if you have a network like PCORnet where all of that data is standardized and in an accessible form, and the opportunity also to, with consent, go back to patients and conduct not only observations but interventional trials. Think about what that all can do for you if you attach to that the electronic health record; genomic information, such as DNA sequence; proteomic information; and very complicated and data-rich databases such as imaging. It's a synthesis of all those things not previously possible that gets us excited but also makes us realize what a challenging problem we have ahead of us just because of the size and the complexity of those databases.

Dr. Joe Selby: Dwayne, if I could just -- this is Joe again -- if I could just add a bit to what Francis said. Historically, we have moved ahead, oftentimes through the conduct and publication of randomized trials, but the truth is, we've had to skimp on these, over time, in terms of their size. And that's just because clinical trials have been so expensive to conduct. And we have paid very little attention in these trials because we couldn't to whether, you know, one treatment is better than another in 100 percent of patients, or whether, perhaps, one treatment works better for some patients but another treatment works better for others. Or a treatment works for some patients and simply does not work for others.

We know, particularly from oncology, that that's true for genetic reasons, but there are also other reasons -- socioeconomic reasons, reasons of age and co-morbidity. So large data, big data, both in our trials, they need to be bigger but also in our observation studies begins to get us to the point where we can legitimately ask the question -- what works better for whom? Some people call this "personalized medicine," but the truth, whether we are talking personalized medicine or just finding the patient subgroups in whom a treatment does or doesn't work is that takes a lot more data than we've had at our fingertips to base our decisions on in the past.

Big data, if it's collected efficiently, if consents and health system involvement are managed well, really open the door to what many people call "personalized approaches to medical care."

Dwayne Spradlin: Thank you very much. So for the next question, I want to play off the word "personalized" there a bit, I guess, and direct this to Dr. Bourne. Dr. Bourne, we have, along with vast volumes of data now available that could be used for research and these studies, we also have new kinds of data -- different data streams coming from tidbits, personal devices, consumer devices and also stunning volumes in new applications coming from, you know, FDA-approved medical devices, you know, glucose meters, and so forth.

This creates sort of an interesting calculus where data streams can now be generated not traditionally in the four walls of the research institution or the clinical facility. How do you think about these new data sets? How they fit into the equation? And how will we manage it as part of this broader fabric of health data that we use for research and make available for study?
Dr. Philip Bourne: Yes, I mean, it's a good question, and, I mean, that influx of that data is clearly perturbing the system. And I actually -- it's a strange thing to say, but it makes me think not so much about the data aspect of it but almost the sociology of it. Perhaps for the first time, health care is becoming patient-centric. It sounds like a silly thing to say, but, really, in many ways it empowers the patient with data of their own. And how they use that data, I think, is, going forward, is going to present a lot of opportunity and also, of course, a lot of challenges.

But they are much more in control of their situation, and the learning and know how to use that information effectively. And you combine that with access to their electronic health record, it creates a lot of challenges for the health care system.

I just know, from my experience prior to coming to the NIH, where I would teach pharmacy students who, often, when they graduate, the first line of defense, if you like, to the health care system between the patient and the health care providers, that we've got to the point where we're actually -- they don't dispense medications anymore as much as they dispense information.

And this whole -- how all this fits together is, I think, the really fun part of what can happen, going forward.

Dwayne Spradlin: Great, thank you. Okay, so this next question -- we've got a few forms of the question being asked by the audience. I'm going to try to consolidate it down, but it plays off this word, Dr. Bourne, that you used -- "control." And I'll make this question open for all the panelists.

We are creating vast pools of data, which potentially create large profit motives for organizations and, you know, obviously, some of the potential for use and abuse that's been talked about in some of the privacy discussion. But how do you think about the use of this data and how it's made available to device manufacturers, pharmaceutical companies, and others that have not traditionally had access to this kind of data? How should we be thinking about that?

Dr. Francis Collins: This is Francis Collins, I'll take first crack. Clearly, there's a tension here, and there will be, no matter how this plays out between a position taken by some that all such data should be completely private, should not be seen by anyone even in an aggregated form, versus others, who would argue the goal here is to try to catalyze medical advances that are going to help people. And, frankly, in order for that to occur, either you need both the academic world and the private sector to be able to take advantage of new knowledge to accelerate those kinds of clinical benefits that we're all waiting for.

So it's going to be important to try to balance those competing perspectives. In a perfect world, you would want to be in a circumstance where our patients do have complete control over every aspect of their medical care data and have the ability to decide who they wish to share it with and who they don't.

In PCORI -- in the PCORnet, the PPRNs are going to have a lot to say about this, because these are patient-powered research networks that are very much driven by people who have thought a lot about these issues. There are organizations like Private Access. If people haven't looked at their website, it's sort of an interesting example of a framework
in which this kind of patient control can be actually instantiated and might be worth having a look.

But I also think that if you really believe in the advances in medical research that are going to benefit the public at large, there need to be examples where aggregated data, which really puts nobody at any personal risk, can be seen by those who have a reason to use it for benefit. And, obviously, those boundaries between these various categories of use are not as bright and shiny as they might be, but I think they are negotiable.

Again, I would want to reassure anybody who is listening to the webinar that these issues are front and center for all of us involved in this new era of big data clinical research, and that there will be no perfect solutions, but I think they will be good solutions.

Dr. Joe Selby: Dwayne, I'll just add a bit from PCORnet's perspective. I can't speak yet to the final policies about data sharing and open data of PCORnet because that's part of what will be worked out over the next 15 months. But what I can say is that decisions about who data will be shared with and for what purposes aren't going to be left in the hands of the networks that participate -- the patient-powered networks and the clinical data research networks, and both will have substantial input into governance and decision-making from the patients involved. In fact, I think networks would be ill-advised to join a project that the patient representatives didn't support.

As I said, we put industry on our steering committee precisely because we believe that there are going to be many questions where patients and industry agree that this is an important piece of research that needs to be done. So our solution is to leave the decisions about the uses of the data to the networks and to the patients and the health system and the clinicians in those networks.

Dwayne Spradlin: Terrific. Okay, so the next-to-the-last question that I'll put to the group, again, is consolidating a few questions from the audience. But it's really taking the point of view now of we've got physicians at the point of care, there are, you know, many different kinds of data now available and, through these programs, increasingly more will be available for research. How do we best equip clinicians with what they need to take advantage of these vast pools of data and actually impact the patient and their health? How do we turn on a system that actually enables the doctor, the clinician, the nurse, to take advantage of all of these data sources?

Dr. Philip Bourne: Yes, I could say a couple of things about this. I mean, I think, clearly, training, going forward, is absolutely critical in this regard. And it's something I think we need to address within the academic medical center arena that we're perhaps not doing as much as we can. And, certainly, there are -- with some of the training programs that we are currently thinking about -- we are actually moving in that direction.

And then I think the other aspect is we mustn't lose sight of is usability. That underlying usability, of course, implies quality data that assuming that at some level, then, usability can take a significant amount of resource to provide. But it's absolutely critical if this data is going to be interpreted correctly and easily used. And so I think that becomes somewhat of an engineering task that can't be underestimated.

Dr. Joe Selby: I'll just add that I think, in some ways, the clinicians are the forgotten stakeholders in all of this. On the one hand, we count on them to put data into the electronic health record in ways that we can -- with a level of quality that allows us to use the data for research.
What they get out of it other than more work remains to be seen. One thing that we don't want to do is foist new data solutions on them before they've really been proven, and I think that often happens these days.

So my sense is that involving clinicians in asking the questions and in generating the evidence is as important as involving patients or systems. But I think we need to pay particular attention to this. We don't want to leave them out of the mix, and we don't want to overwhelm them. We want to provide them with genuinely useful new tools that they understand and actually have expressed a need for.

Dwayne Spradlin: Perfect, thank you. So for our final question, I'm actually going to ask for a prediction from each of our panelists. We have very exciting and intriguing tools. We are creating the ability to collaborate around health data research, patient records, on a scale we've never done before. But if you were to fast-forward, say, five years, what would your prediction be? And why don't we start -- why don't we go in reverse order. We'll start with you, Dr. Bourne.

Dr. Philip Bourne: I would -- I don't know about a prediction, it's more of a hope -- that we would see, as a result of the big data revolution, whatever that means, but certainly it means translation, and it means integration of data in a way that Francis described, particularly. That that would lead to a major breakthrough. We will see major breakthroughs in our treatment of disease as a result of this kind of activity.

Dwayne Spradlin: Wonderful. Dr. Collins?

Dr. Francis Collins: I would agree with that. I think we will, because of the ability to mine this kind of data and to do so with relatively low cost and high throughput we will have a much better handle on what works and what doesn't work in the real world enabling, therefore, both prevention and treatment and cure to be carried out in a fashion that's based on a much broader evidence base than what we currently have access to with benefits to millions of patients. That would be my hope and my dream. Five years is a pretty short time. If you'd said 10 I'd be a little more confident, but let's go for five.

Dwayne Spradlin: Great. Okay, Dr. Selby?

Dr. Joe Selby: Well, I'll agree with Francis that five years is a short time as things evolve. But I would predict that in five years we'll have examples of big studies that look different than studies that have been mounted in the past because patients or clinicians have had a hand in crafting them. So we will see a subtle beginning of a shift, but there will be examples we can point to where research is actually beginning to focus more on the questions patients and clinicians have.

Secondly, I'd like to think that big data will very gradually begin making inroads in predicting optimal treatment choices. So I think there is an immense amount of room for using data to help guide individual decisions about whether a treatment makes sense and which treatment makes more sense. But that is a big task. People have been working on it for years, and there aren't many examples yet, but I think we'll see more progress in that area.

And, lastly, I think we will actually see health care delivery systems beginning to believe in data and in -- actually, in research, as a way to improve their care delivery.
Dwayne Spradlin: Terrific. Well, thank you, everyone. We are now at the end of our time today. I want to thank Dr. Selby, Collins, and Bourne again. Our panelists today were outstanding, and I'm sure you'll agree. These are incredibly important topics, so we have a tremendous amount of work ahead of us. Collaborating and sharing through activities like these are an integral part of coordinating all of our work, some vital if we're really to put health data to work, as we must.

If you have any questions or comments, please do not hesitate to send them to info@healthdataconsortium.org or through Twitter at our handle, hbconsortium. In some cases, we'll be able to get those to the presenters and potentially help you get responses. Also, please note that the archive and transcript of this webinar will be available in a few days at www.pcori.org and on the HDC website, www.healthdataconsortium.org, as well, very shortly.

Finally, as some of you may know, the Health Data Consortium is the convener of Health Datapalooza 2014, which is coming up in a few days in Washington, D.C., June 1st through 3rd at the Marriott Wardman, to be precise.

I think all three panelists will be in attendance on panels and in different roles and, in fact, Dr. Collins will be giving a keynote on Tuesday. We'd love for you to be able to join us there.

With that, I want to thank the panelists again, all of you that have listened in and attended and asked questions here today, be well. This webinar is now concluded, thank you.