FINISHED FILE

PUBLIC HEALTH DATA SCIENCE: THE NEXT DECADE
FEBRUARY 29, 2024
1:00 P.M. ET

Services provided by:
Caption First, Inc.
Monument, CO 80132
719-481-9835
www.captionfirst.com

>> SANDRO GALEA: Good afternoon, good evening, good morning, wherever you are. My name is Sandro Galea. I have the privilege of serving as Dean of the Boston University School of Public Health. And on behalf of our school, welcome to today's Public Health Conversation. These conversations are meant as spaces where we come together to discuss the ideas that shape a Heather world. Through a process of open discussion, debate, and the generative exchange of ideas, we aim to sharpen our approach to building such a world. Guided by our speakers, we work towards a deeper understanding of what matters most to the creation of healthy populations. Thank you, everyone, for joining today's conversation. In particular, thank you to the Dean's Office and the Communications team, without whose efforts these conversations would not take place. Thank you to the co-hosts, the BUSPH Population Health Data Science program and the Boston University Faculty of Computing and Data Science.

We are here today to talk about data. The evolution of data science and the evolution of public health are deeply intertwined. How we engage with the increasingly vast reserves of information at our fingertips will have profound implications for how we shape a healthier world. The rise of population health science and Population Health Data Science represent enormous opportunities we can leverage towards a better future and more effective pursuit of health today.

I'm now delighted to introduce today's moderator, Dr. Debbie Cheng, Professor of Biostatistics at our school and the Director of the School of Public Health's Population Health Data Science program. Her research focuses on applied statistics and the design and analysis of clinical trials. She is the associate Director of the Providence/Boston center for AIDS Research and co-Director of the Biostatistics Core for the CFAR. Dr. Cheng is also the director of the biostatistics and data management core for the International URBAN ARCH Center, which aims to examine the impact of alcohol on multiple aspects of the TB disease continuum. On a personal note, I have always loved learning from her. Over to you.

>> DEBBIE CHENG: Thank you, Dean Galea for the kind introduction. Welcome, everyone. It is such a pleasure for me to

be moderating today's discussion. To get us started, I'd like to begin by introducing our very impressive panel of speakers today.

First, we will hear from melody Goodman. Dr. Goodman serves as Senior Executive Vice Dean and Professor of Biostatistics at NYU's School of Global Public Health. Dr. Goodman's research seeks to develop a more rigorous understanding of the social risk factors that contribute to health inequities in urban areas with the goal of developing culturally competent, region-specific solutions through collaborative activities with community members, community-based organizations, faith-based organizations, and other community health stakeholders.

Next, we will turn to Yulin Hswen, an Assistant Professor in the Department of Epidemiology and Biostatistics at the Bakar Computational Health Institute at the University of California San Francisco and faculty in the joint Computational Precision Health Program with UC Berkeley. Dr. Hswen's research is at the intersection between the digital environment and society, where she focuses on using new artificial intelligence and machine learning methods to uncover social patterns of disease and to develop unbiased and fair systems of health.

Then we will hear from Michael Kosorok. Dr. Kosorok is the W.R. Kenan, Jr., Distinguished Professor of Biostatistics and Professor of Statistics and Operations Research at UNC-Chapel Hill. His research expertise is in biostatistics, data science, machine learning, artificial intelligence and precision medicine. He has written a major text on theoretical foundations of these and related areas in biostatistics, as well as co-edited a research monograph on dynamic treatment regimes and precision medicine.

Next, we will turn to Nathan Lo, Assistant Professor of Infectious Diseases at Stanford University. Dr. Lo's research group studies the transmission of infectious diseases and the impact of public health strategies with the ultimate goal of informing public health policy. His research blends diverse computational methodologies, including tools of epidemiology, modeling, pathogen genomics, and policy analysis.

And then, finally, we will hear from Daniela Witten, a professor of statistics and biostatistics at the University of Washington and the Dorothy Gilford Endowed Chair in Mathematical Statistics. She develops statistical machine learning methods for high-dimensional data with a focus on unsupervised learning. Dr. Witten is the co-author of the textbook "Introduction to Statistical Learning." She also serves as Joint Editor of the Journal of the Royal Statistical Society, Series B.

We are so fortunate to have such a highly esteemed panel of experts here today, and I know we're all looking forward to hearing from them. So, without further delay, I'm delighted to turn things over to our first speaker, Dr. Melody Goodman. Melody, over to you now.

>> MELODY GOODMAN: Thank you. Hopefully, you can see my screen. So, I just wanted to talk about where I think some important things for us to think about in the next decade. And this paper came out about addressing health and health care disparities: The role of a diverse workforce and the social determinants of health. And as someone who spent their career really focused on the social determinants of health, I really started to also work on trying to create a diverse workforce.

In 1999, the Founding Dean of NYU School of Global Public Health, Dr. Charlotte Hilton wrote a Seminole paper called "The Shape of Our River" and it looked at the racial and ethnic diversity of faculty, students, and graduates in schools of public health.

In 2019, we did a 20-year update to that paper, looking at the racial composition of member institutions and the associations of schools and programs in public health. And in doing this work, I got really interested, because I was like, I wonder how diverse the quantitative side of public health is. And so, we looked at, particularly, the racial and ethnic diversity amongst students, graduates, and faculty in biostatistics and epidemiology.

And what we're seeing is that we're not as diverse, probably, as we would like to be. Although diversity's increasing, I think we definitely have some ways to go.

And so, in terms of STEM, if you think about STEM, there's been this idea that there is a leaky pipeline, and there's data to support this idea, that there's a leaky pipeline leading into disciplines such as biostatistics and data science. However, my hypothesis is that the pipeline is not leaky; it's that the pathways are different. And if we make it really easy for certain types of people to come into our field and harder for other types of people to come into our field, we're going to see more of the people that it's easy for them to come in and fewer of those who it's harder to come in, and it's our job to create pathways so different types of people can enter the field with relative ease and supports.

And so, a lot of us have been working on pathway programs, and I think these are important because, mainly, they give participants three things: A window; a window that allows participants to get a glimpse at a career and what people in that field do; a mirror that allows participants to see their reflections -- someone in the field that looks like them; and the great ones offer an open sliding glass door where participants can walk through and get hands-on experience in that space.

One such program I developed in 2023 was a partnership between NYU School of Public Health and the John J. School of Justice, part of our New York City system. The students had 20 hours of foundational and introductory training geared towards their social-emotional well-being and skill development at John Jay and then spent four weeks in June in an intensive training force at the NYU School of Global Public Health, and they were all in. It was a really exciting group of students to teach. At John Jay, they learned about data exploration and data management, formulating research questions and hypotheses, determining validity and reliability, and finding data sources. They learned to use Excel to sort data and conduct data entry and do some coding. And they also had some sessions on setting goals, developing a growth mindset, and navigating imposter syndrome.

When they came to NYU, our goals were for them to gain both conceptual and practical skills in data management, data visualization, and data analysis, to provide them with an introduction to public health and exposure to careers at the intersection of public health and criminal justice, and for them to learn how to read, understand, create, and communicate quantitative data as information.

Just want to share a short video. Hopefully, it will play for me.

(Music)

>> I am a statistic. I am the one out of three who will go to college.

>> I am the three out of four who don't do drugs.

>> I am the five out of nine who have a job.

>> I am the seven out of eight who is not a teenage father.

>> I am the 11 out of 12 who won't drop out of high school.

>> I have a purpose, and that's a fact I'm proud of.

>> MELODY GOODMAN: So, I love that video, and we show it to our students, because a lot of times, we hear really negative statistics about black and brown men, but there's lots of positive statistics out there about them as well. That's one of the things they learn in our course is how to show, demonstrate some of those positive statistics.

They learned a lot in terms of introduction to public health, structural racism and community health, research methods, biostatistics, epidemiology. We had eight sessions focused on data literacy that I taught them, and then we also had some from our Department of Health come in and talk about data sources, and a professor from John Jay talked about the intersection of health and law.

In addition, they had professional and personal development courses which focused on public speaking, storytelling, resume and cover letter writing, job search and networking techniques. They were trained to administer Naloxone and they also had a session on time management and professionalism.

And so, part of the data literacy training, they have a really eight-session intensive course where they learn both data and are introduced to Tableau, just getting their hands on some software tools. And our culminating experience was a Datathon, where they had a day and had a challenge problem and spent the day finding a solution to the challenge problem. The challenge problem came from an organization in New York City, the Drug Policy Alliance, and they were interested in both the best attitudes that reduced harms of both drug use and drug prohibition and promote the sovereignty of individuals over their minds and bodies.

A lot has happened since we ran our programs this summer, including the Supreme Court decision around race-based admissions, and we have written a commentary where we think the implications are for public health, but we also know that some schools are taking this and looking at it just in terms of admissions, and some schools are looking at this decision more broadly, and we think that there are some real implications there.

I want to close up by saying that I think the last thing that our field needs to think about is the ethics, that as biostatisticians, that we are narrators and that our data do not speak for themselves, and it is important for us to be ethical in the stories that we choose to tell and how we choose to tell them. Thank you.

>> DEBBIE CHENG: Thank you, Dr. Goodman. Thank you. Thank you for that great presentation. Emerging leader sounds like such a wonderful program.

At this time, I will turn it over to our next presenter, Dr. Yulin Hswen. Yulin, please take it away.

>> YULIN HSWEN: Great. Thank you so much. Beautiful presentation, Melody, so thank you very much. Okay.

So, I'm going to open with this quote, which is, "With artificial intelligence, we are summoning the demon." And I think that that's kind of just the future that we're heading is really AI, and the questions around whether or not we are opening Pandora's box and whether or not AI can do good or it can do evil.

I think in terms of the evolution of data science in the context of public health is, I think we're really in an era where data is going to be informing our decision-making. And I think that is being seen through the use of electronic health record data, cell phone data, social media data, in terms of identifying risks, informing interventions, and you know, even if we don't know it, large tech companies are shifting our behaviors and giving us nudges.

And I think the real discussion around, should be around the ethics of this and whether or not we should just be focusing on data and whether or not there needs to be a human aspect to the decision-making, which leads me to my next point, which is this emergence of generative AI and large language models. So, for example, ChatGPT. At least at UCSF, there's a lot of work that's being conducted around the use of ChatGPT, you know, potentially providing advice to the public or to patients, or the use of it in taking clinical notes, for instance. And there are debates around whether or not, you know, this type of technology can increase access, or whether or not we're just lowering the standards of quality and care for less-resourced communities and populations. So, the larger question around whether or not it's going to reduce disparities or actually widen them, because we are going to become almost a little bit complacent in terms of providing the same level of quality of care to certain populations if we have this kind of artificial intelligence tool that we give out.

I think the new types of methods that we're going to use to shape and inform population health -- I'm beginning to work on a type of method called synthetic controls. I think more and more, we're going to have to use this big data to understand research questions and causal inferences. Obviously, something that is needed for these type of data sets. So, these synthetic controls are a way to, again, mimic our randomized control trial where we generate a weighted control group and are able to potentially compare it to a treatment group so that they resemble similar kind of, again, starting pathways and seeing what happens when the treatment is provided.

I think that, also, we need to start looking at more econometric methods. I think we need to start thinking bigger and measuring policy, social interventions, and evaluations that are iterative, like interrupted time series, difference in difference. And I think we need to start measuring these type of interventions and evaluating them and then being able to iterate on them, if they are not being effective in real time.

I think the other really large area that's happening is the natural language processing and type of sentiment analysis. I think there are going to be lots of new kind of tools that just focus on language. You know, that type of interaction with AI. It's managing large amounts of text and deciphering and decoding a large amount of text, both from these large language models, but then also from, you know, responses from the public or from the patients, being able to identify what is most important and what is needed and bringing in the public voice.

So, the other quote that I have is "The Terminator would never stop. It would never leave, and it would never hurt him, never shout at him or get drunk and hit him, or say it was too busy to spend time with him. It would always be there, and it would die to protect him."

So, this idea around the AI tools do not get tired, and there is something that is occurring, which is clinician, and just overall burnout across all populations, in terms of kind of the workforce.

And so, there's been these studies that are looking at comparing these artificial kind of intelligence technologies and whether or not they are as empathetic as humans are. And it turns out, in at least one of the studies out in the Journal of Internal Medicine, it was rated that the responses from ChatGPT were rated more empathetic than physician responses.

The other big area that I mentioned kind of with, like, language and so forth, is these new type of kind of AI models and methods being able to translate brain waves into kind of speech. And so, these deep kind of learning models are being able to essentially read our minds.

And so, I think, again, the other question -- and Melody kind of touched on that as well -- is the ethics kind of, of the data. You know, what happens when you are able to decode what people are thinking? And yes, it's, again, for the benefit of being able to potentially give people speech back, for those who have, for instance, lost their speech with a stroke or so forth, but what are the ethics behind all of this data that we have that's very personal?

Again, neural decoding, large language models. Again, I think language is going to be a very big part within the next decade or so, and using generative AI -- I know a lot of people, offline discussions, have been telling me how much they use AI to produce information, produce policy briefs, and so forth. So, is this already being used to kind of generate policy and practice.

And then, kind of lastly is this education and training and the kind of next wave of students and learners in the future, whether or not the use of generative AI is something that will be, you know, used instead. So, for instance, ChatGPT passing the USMLE, and what does that mean for medical education, for instance, and are students kind of allowed to use that? So, just kind of leaving you, overall, with another quote -- "If a machine, a Terminator, can learn the value of human life, maybe we can, too." I think I'm really kind of thinking about, kind of overall, is -- you know, we focus so much on AI, but I think we need to also start making sure that the data that we have is unbiased, and that's by making sure humans are unbiased. But we still kind of need these, like, real-world testing diverse data sets, and I think we need to really start doing experiments on testing what we call automation bias in AI for decision making, and just the effectiveness of these new type of AI tools in making sure that they're providing the advice and care that is necessary across all these social determinants of health, before that we deploy them. Okay, thank you very much.

>> DEBBIE CHENG: Thanks, Dr. Hswen. Thank you. That was excellent. Next up, we have Dr. Michael Kosorok. Michael, over to you.

>> MICHAEL KOSOROK: Thank you. And Yulin, that was a very interesting and powerful talk. Thank you. So, I'll be talking a little bit about some of the artificial intelligence work and related work that's done in my lab at University of North Carolina Chapel Hill, the precision artificial intelligence research lab. This is a photo of the subset of our members. We have about 30 people in our lab. It's mostly PhD students, but we have high school and undergraduate students as well, postdoctoral fellows, K-scholars, and they come from diverse health-related departments and areas. And our focus is on using artificial intelligence and machine learning to advance precision health, as well as other machine learning and AI challenges in biomedicine.

So, I want to go through a couple of examples briefly of some of the work we've done and some of the questions it raises and some of the questions, hopefully, it answers. First is work on precision treatment of chronic lower back pain. The goal here is to -- this is related -- this is part of the BackPack initiative which falls under the HEAL initiative, some of you may have heard of. The goal is to develop a data-driven treatment algorithm to optimize outcome for each patient specifically. We wanted to develop an algorithm that a patient presenting with chronic lower back pain, and based on their biomarker information -- meaning anything about them, demographic or otherwise -- based on that information, we assign an initial treatment, and after 12 weeks, depending on how they respond, we will assign a new treatment to them or a combination of treatment or switch, whatever appears to be needed. Then at the end of 24 months, it's a hope that they will achieve a maximum reduction in pain.

And so, we use a sequential multiple assignment randomized trial, or SMART trial design. This enables efficient discovery of the unknown treatment regime or algorithm, I mentioned before, where our first aim is to estimate the optimal regime that minimizes pain at 24 weeks and the second is a dynamic treatment regime that optimizes depending on the patient's priority for their outcome. So, they're, in addition to pain, there is quality of sleep, clarity of thinking, enjoyment of life, and a number of other ones that may be important for some patients.

And the machine learning tool it uses off-policy reinforcement learning.

Here is a schematic of the design, which is admittedly complicated looking. I do want to say that this was put together over months in collaboration with a team, a consortium of collaborators from across the United States, mostly consisting of clinicians, but also biostatisticians. Biostatisticians took primary responsibility in the design of this study with support and guidance and insight from our clinical collaborators.

We used four treatments, initial randomization that established treatments. What is not known is for whom they work best. So, the idea here is to link these biomarkers I mentioned and other demographic information so that we can actually know how to assign treatment initially. And then, after 12 weeks, the patients are re-randomized, depending on their response to the previous treatment, and it's either to switching the drug or the treatment to, could be also augmenting, and then they're followed for 24 weeks.

As far as I know, this is the largest SMART design for clinical decision making. There are many for behavior and other types of decision making, but this is as far as assigning treatment, this is the largest we know of. And it has over 900 patients enrolled, and we are very much on target within a few months to have well over 600 completers of this study.

Okay, so, see just want to point that out, that that's something that we've been doing, we're working on, and is AI related, and we hope to have a significant impact on quality of life for these patients.

I'm going to switch topics. I'm kind of moving around a bit here. I also want to talk about using machine learning, using causal inference-based techniques to be able to understand better some of the causes of health disparities. I want to talk about a particular project that we completed, published in 2022, in diabetes care, looking at health disparities in type I diabetes among youth and young adults.

This is a paragraph of many of our wonderful co-authors on this paper. We used the data from the SEARCH for Diabetes in Youth Study. And I want to say, the basic premise here is as follows. First off, we note that in this population, Hba1C, which is a marker for severity of type I diabetes -- the larger the number, the worse; the smaller the better -- for our non-white subgroup, the Hba1C is about 9.2, and for the white patients, that's 8.2. That is a very significant health disparity.

And what we looked at was, this is longitudinal data. We use off-policy reinforcement learning. And what we do is we look at what would happen to this same population, had we, contrary to fact, given the non-white patients the same distribution of treatments available to the white patients, stratified by things like age and other factors that are demographic. And with that, what kind of difference would that make?

We found that this addressed 40% of the disparity. It was very statistically significant, suggesting that this simply reassigning the same -- ensuring that the same treatment options are available can reduce some of the health disparity issues dramatically. We know that there are other causes of health disparities, but I think this is an important rigorous analysis, first of its type to begin to look at specific things that could be done.

Okay, another unrelated project is looking -- but related in some ways -- is looking at some technique we developed for automated gestational age prediction. This is a global health issue. Our focus is looking to help in underresearched locations where communities do not have access to physicians, nor expensive ultrasound equipment, and gestational age is very, very important in health of delivery of babies. And we want to make sure that more individuals can have an accurate idea of the age of their child.

And so, we developed a deep neural net algorithm that uses a similar technology, transformative technology that's used in ChatGPT, but modified for our purposes here. And the accuracy of this in our study on the test set is slightly better than expert. We had (?) in this, so that helped us.

One of the things I wanted to mention, our goal is to deploy this in underresearched location with a set of a multi -- you know, $100,000 ultrasound, using a handheld, $1,000 or less ultrasound, and instead of an expert, have an iPad or iPhone to do the calculation, prediction of the age.

And what we've also done is we have just completed a field testing, field study, to evaluate this in the settings where we want it to be working well.

A question that this raises: How should we evaluate and deploy new AI technology for health? Especially in this case, as the results of this may be used, and should be used, probably, for decision making, when there is not a physician, a trained person available. I'm not saying we shouldn't strive to have trained physicians available everywhere, but in some underresourced locations, that's not feasible, and this will provide greater safety and health to these individuals than not doing anything.

We also feel it's important to work on performance guarantees for these models, just for reassurance. We don't have a lot for these, but we can evaluate them. And I mentioned our field trial. It shows that right now it's under review. The preliminary results are very positive. Is it safer to deploy this tool than not is the question? We think the answer's yes, but we think those are important questions to ask.

What long-term responsibilities do we have with such AI tools? Would a phase IV follow-up study-type model be good to incorporate, so that we're always making sure that this is safe and usable throughout its utilization, and adapting as needed. Okay.

So, my goal for these three different projects is just to show that AI is being used in many different ways to address many questions. And I want to say that AI is here to stay. It will only get more interesting. I'm hoping that's a positive type of interesting, with or without our engagement. I think as quantitative data scientists, and as everybody who's interested, this is very important for us to be involved in.

I also want to point out that data-driven personalized and population level realtime decision support is very important. And one thing about biostatistics is we're the science of biomedical science. Our job is to sort of help people know how to conduct good, data-driven science. And I think it's important for us to lead from the front and not always from behind, meaning that in addition to helping people with the questions they have, we should help ask those questions and develop methodology to answer those questions, AI tool. They'll answer questions that maybe people aren't thinking about asking right now.

And I do feel that every -- we need to include people from all groups in all steps of the process here -- in the research, in the education, in other steps. And I think all of us would agree that for this to be successful, we need to make sure that humans -- not just AI -- are good at critical thinking and compassion. And so, thank you.

>> DEBBIE CHENG: Thank you, Dr. Kosorok, that is fascinating work. We are already having questions coming in from the audience in response to the presentations so far. That is great. I want to encourage the audience to please keep those questions coming, and we will address as many as we can during

the Q&A session following these presentations.

For now, we will turn to Dr. Nathan Lo. Dr. Lo, please take it away.

>> NATHAN LO: Thanks for the kind invitation to present today. I will focus on novel data and predictive analytics in outbreak response, and generally, control of infectious diseases. I have no commercial disclosures.

So, in 1928, Alexander Fleming discovered penicillin. And at the time, people thought, you know, the age of infectious diseases was over. But I think over time, we've been continuously humbled to recognize that infectious diseases continue to remain a threat. Of course, that was on full display during the COVID-19 pandemic. But even since the emergence of SARS-CoV-2, we have had re-emerging infections, such as Mpox and somewhat unexpected resurgence of prior and known threats like RSV, further elucidating and suggesting these infectious diseases continue to remain a threat.

So, the topic of my talk today is, what are the recent advances in data science, specifically those made during the pandemic, to support outbreak response and control of infectious diseases? And for my talk today, I want to highlight two aspects I find interesting. The first is cell phone app-based contact tracing software, and the second is generally forecasting and modeling tools for public health action.

So, first, app-based contact tracing. So, you're likely familiar with the cell phone software that told you during the pandemic when you were likely in contact with someone with COVID-19. Now, here in Boston, the predominant software was MassNotify. In California, where I reside, that was CA Notify. But perhaps the most widely recognized and implemented was the NHS COVID-19 app in the United Kingdom, their official contact tracing app.

As many of you know, the way this contact tracing app works is people opt in, and then over time, your cell phone essentially communicates with people as you walk by and spend time with them. And when one of those people is diagnosed with a SARS-CoV-2 infection, if you meet some risk score for an exposure, an alert is then sent to all the contacts of that person.

Now, every app has their own, you know, score and risk threshold for defining what is a high-risk exposure. This is the risk score for the NHS app in the UK, which I just, you know, found interesting to share. And it essentially has three components. The first is the proximity score, which is like how close you were to someone. The second is the duration. And the third is the infectiousness of the COVID case based on timing. And this is like the first time that an app like this was rolled out for controlling infectious diseases at this scale. This is certainly a big deal in terms of data system and structure and data analysis, but it was the first time this was on full display for control of an infectious disease.

And in terms of the contributions of the tool, I kind of think of them as two-fold. The first is pretty obvious. It's like a public health tool to reduce transmission. But the second point that I find really fascinating that I think is underrecognized is that it's a really unique scientific data collection platform to allow us to learn key and time-sensitive, and perhaps evolving epidemiologic features of a particular pathogen.

So, a question that you might have right now is, did this work? Was it an effective public health tool? And for that, we can turn to data from the United Kingdom and the NHS contact tracing app. This was a paper published a couple years ago, where basically, they looked at the NHS app over a three to four-month period, and essentially, from analysis of that data, they were able to estimate that about 280,000 to 600,000 cases were averted through effective quarantine of contacts. So, this is over a three to four-month period, and this is accounting for imperfect quarantine, delayed notification, et cetera.

But perhaps more poignantly, essentially, for every case that consented to notifying their contacts, at least one case was averted. And so, this is really exciting, because this is a tremendous undertaking and the first kind of, like -- or at least the largest rigorous evaluation of this type of software as a, you know, more advanced form of contact tracing.

Now, what I will say, what is interesting is, this same level of evidence is not necessarily true in the U.S. So, the UK has many advantages here in terms of being a single health care system; the app being integrated within NHS and their testing platform. And so, you know, the same rigorous data isn't necessarily available for all of the other app-based contact tracing softwares. And I think the generalizability of these findings to other locations, and certainly other pathogens, would still be less clear.

But the other feature I find really fascinating is, like, how can we use this data to really learn important, critical questions, such as, what is a high-risk exposure? Like, what is the degree of exposure where truly I'm at risk for getting an infection? And for this, once again, we can turn to the same group from Oxford who looked at 7 million contacts, and from that, was able to, like, really high resolutionly -- from a high resolution standpoint, characterize, what is a high-risk exposure? Looking at the relationship between time of exposure and risk of infection, household versus community, and ultimately, evaluate kind of a composite score, and really use this as key information to kind of more precisely determine things like quarantine. And so, I think this is like an example of a really interesting approach, certainly one that is novel and new for both time-sensitive studies, but also public health measures. But also to note, it's an approach that's most likely to succeed in a country where both the outcome data and the contact tracing data are really integrated in a single system.

And what I'd be excited about is the potential for application across all infectious diseases, such as during the Mpox outbreak, but certainly a lot of advances needed.

And what I will say, of course, is there are many limitations that I could spend the whole time talking about, including privacy considerations, biases in the populations who opt in, low-case ascertainment, and of course, low uptake of the app. And these are all important barriers that merit discussion.

The second area to talk about is forecasting and modeling tools, and this is kind of my area of research, which is developing predictive modeling tools for public health action. Now, these tools are not new, but really, their application to decision making and public health decision making is changing and really evolved during the pandemic. And when I say these predictive modeling tools, I mean those used for forecasting cases and hospitalizations, for vaccine and testing decisions, and other response measures. And these modeling tools range from statistical methods, including a lot of machine learning approaches, but also mechanistic models that are popular in my field and other ensemble and modeling approaches.

So, I'd like to highlight just a few new things in this space. The first is collaborative initiatives for forecasting infectious diseases. Now, a lot of these initiatives, such as for flu, far preceded the COVID pandemic, but I think these collaborative initiatives were really on display during the pandemic. This includes the COVID-19 Forecasting Hub, the Scenario Modeling Hub, to some extent, IHME, and also state public health departments. Essentially, these are initiatives where multiple academic independent groups submit their forecast based on their own, you know, inhouse models, but in a standardized format. And then, these initiatives then allow us to understand what works, what doesn't work, do a really rigorous evaluation of our strengths and limitations, and also just generally lend transparency and more credibility to a lot of these model-based estimates.

The second big change in the pandemic is CDC itself started a Center for Forecasting and Outbreak Analytics. The center, itself, does the work, but also with many collaborating centers, including investigators here at Boston University.

The third key aspect is the integration of novel data streams into models. So, increasingly, we have really important data from genomic surveillance, wastewater surveillance, and mobility data; yet, how to exactly integrate them all together into predictive modeling frameworks for public health decision making still is less clear. A lot of open questions remain, such as, what is the value of these different data sources for different decisions? What are the methodologic advances needed to integrate them together effectively? What are the performance characteristics of these data sources for a decision? And how do we approach basic questions like optimal sampling?

And then, finally, the fourth point that I'll say is, there is a lot of focus increasingly on how to develop modeling tools that can be embedded and integrated in public health agencies. My own work focuses on, my NH-funded work is about developing predictive modeling tools for public health departments and then rigorously evaluating, how are they used, how can we make them better, do they change decisions, and do these decisions causally improve outcomes? And I think really focusing on developing these tools out of the academy and into public health practice is really a key frontier.

So, in summary, these modeling tools, the methodologic advances include integration of novel data, integration within public health agencies, alongside rigorous evaluation; of course, the quality, timeliness, bias of data are critical considerations. And additional considerations include models that ensure equity, address data challenges, and ensure an uptake of models and rigorous evaluation and validation. Thank you very much.

>> DEBBIE CHENG: Thanks, Dr. Lo. Very, very insightful presentation. Thank you.

All right, next up, we have Dr. Daniela Witten. Dr. Witten, over to you now.

>> DANIELA WITTEN: Great! Thank you so much. I'm so excited to be part of this discussion today. I'm a statistician, and the thing that I'm going to talk about is sort of how statistics is affected and how the field has faced new challenges and new opportunities in light of data science.

So, as a statistician, we sort of have this canonical view of how the scientific process works. This is how we write our textbooks and this is how we teach our students. So, we start out with a scientist who has a really great idea to test Seminole hypothesis that we'll call H inaugurate. And she collects data and based on that data she collects, she accepts the hypothesis or rejects it. And then she publishes her results, hopefully, in a really high-impact journal. Okay. So, this is sort of the -- how we imagine statistics would work. But of course, this isn't really what happens.

And in reality, the scientific process in this era that we're living in of large-scale data looks more like this, where maybe somebody collects a whole bunch of data without a specific, detailed statistical question in mind. Maybe the person, the woman collecting the data -- this example, the scientist -- might be like, oh, yeah, I want to know, like, how this pathway interacts with this disease, but it's not like a specific scientific question. So, then, she might look at her data, develop a hypothesis from investigating that data, test that null hypothesis and either reject or fail it or reject the hypothesis. But she's not done. Now she'll iterate, where using the results of that test, she might explore the data some more and keep on looking and keep on trying different things again and again, until finally, she's ready to publish her results. But furthermore, in real life, there's a little asterisk next to publishing results, because, of course, we only publish positive results.

So, basically, the scientific process is quite different in real life than it is in our textbooks, and I just want to emphasize that these differences are kind of -- there's sort of two differences. One is sort of a new difference, which is the fact that with larger-scale data, we're more and more tempted to sort of iteratively refine the questions we're asking and sort of re-analyze the data again and again using refined questions. So, that's sort of a relatively new phenomenon, I think, relative to maybe the way that science used to be done.

But the phenomenon where we tend to only publish positive results, that's been true forever. And these are both really big statistical issues.

So, why are these issues? Well, if we test our null hypotheses and we look for significance level at some number alpha, so for example, we could set alpha equals 0.05, meaning we will reject the analysis if the P value is less than that, and if you want to use a different threshold for alpha, go ahead, no problem. Then this part of the process, where we're kind of like iteratively refining the question that we're asking, using repeated peeks at the data, we can refer to that as double dipping. And essentially -- and it's double dipping because we're iteratively dipping into our data again and again and the problem is it will cause us to reject our null hypothesis time and time again, so we like to reject that no more than 5% of the time with 0.05, but if we're iteratively looking at the data to choose a null process, we might reject the null much more than 5% of the time and that's a problem because we lose the guarantees of the results. Essentially, the statistics we do become meaningless.

The other part of the problem here, which is, of course, a journal is not going to be terribly interested in publishing a negative result. Only positive results will be published, meaning far more than 5% of published findings are going to be false, even if we reject our null hypotheses at null 0.05. So, these are two issues, one's been around a long time and one is relatively new. And I should mention, this one that's been around for a long time, like this paper dates back to 2005, so almost 20 years old, why most published research findings are false. So, the idea that there are a lot of findings is nothing new and dates back far before 2005.

As statisticians, what are the solutions? One solution from my perspective, it's like pretending we live in a textbook. So, we don't live in a textbook world, but we can pretend that we do by splitting our data into two independent sets -- a training set and a test set, where we then use our training set to develop a hypothesis and then the test set to test it, so that we avoid double dipping. And so, I think about this as pretending that we live in a textbook. And it often can be a really good solution. This is known as sample splitting. So, if you do cross validation on new data or whatever, you're doing sample splitting so that you can, on your test data, pretend that you live in a textbook.

Sample splitting, there's a lot of cases where it doesn't apply, for example, if you have correlations among your observations or if you don't have enough replicates. And my group has developed a new group called data thinning to do the test sets without splitting up the sample. This is very new work that we've been working on.

So, I describe sample splitting and data thinning is pretending that we live in a textbook. The other thing we could do is don't pretend we live in a textbook, and instead, just account for the fact that we live in the real world. And the way to do that is when we test our hypotheses, we should account for the fact that we double dipped our data. We should account for the fact that we selected a particular null hypothesis by looking at the data. This falls under a statistical framework called conditional selective inference that has been around for almost ten years now, which provides a really very interesting way to account for some of these issues that arise in the context of live-scale data analyses.

So, I just want to conclude by saying, you know, we have a problem in our statistical understanding of large-scale datas, which is that a lot of the existing and old tools don't apply, but we are developing solutions. So, this is an exciting time both to be selecting the data and to be analyzing it. Thank you so much.

>> DEBBIE CHENG: Thank you, Dr. Witten. That was terrific and a great topic. All right, so, that was our final presentation. Thank you, again, to all our speakers for the wonderful presentations. I think it's clear our panelists have an amazing breadth of expertise and knowledge.

We'll be moving at this point to the Q&A portion of today's event, so I'd like to ask all of our speakers to please turn your cameras and microphones on at this time. Thank you.

All right, we have several questions that have come in from the audience. That is terrific. And the audience can continue submitting questions using the Q&A feature at the bottom of your screen. I do have a couple of questions of my own that I'd like to begin with, if I may, just to help get things started.

All right, so, to begin, when I think about how the field of public health data science might evolve over the next decade, a question that comes to my mind is around promoting more collaborations among data scientists, because data science is such a broad field. It encompasses many different disciplines -- biostatistics, epidemiology, computer science, engineering, and others. So, several unique and different areas of expertise are under this umbrella. And I'm curious what you all think we as biostatisticians and epidemiologists, what can we be doing and what should we be doing to help foster closer collaborations with disciplines like computer science and engineering, where we do a lot of similar work, but perhaps speak a different language? What can we do to promote more collaboration between various data science disciplines, which I think could help advance all of our work, both research and education. Dr. Lo, I think you have a very unique perspective, having trained in both engineering and epidemiology. I wonder if you'd like to start first with your thoughts on this? And then, perhaps others can chime in as well.

>> NATHAN LO: Oh, yes, yes, thank you. Yes, a very important, but also challenging question. Also being a new faculty member, this is a question I think about a lot in terms of the students that we get to work with and interact with, as well as the faculty that we get to interact with.

I'd say I'd kind of think it in two ways. The first is very narrow, which is in my field of infectious disease, modeling work. I think there where I've seen a lot of exciting aspects is centers that really focus on a particular goal, and the particular training or discipline to achieve that goal comes from very diverse places, but everyone is united within like a single center, allowing for, like, a lot of projects and students and faculty members to work together more seamlessly.

I think on the second point would be like a broader scale across like a university setting, and perhaps someone else might be able to speak more to that. But I think I certainly benefit a lot from spending a lot of time at various centers and conferences and seminars across campus and learning from different disciplines. But I think the other panelists will probably have more insightful things to say, such as like Dr. Goodman, for example.

>> DEBBIE CHENG: Thank you. Dr. Goodman, would you like to chime in?

>> MELODY GOODMAN: Yeah, I think this is an important question, and I don't know that I have the solution, but I do think there's a couple of ways we could foster collaboration. One I think Dr. Lo mentioned is in research centers, but the other way is in our teaching and our training. I think sometimes we can co-teach, and that makes -- even though it's on the academic side, those connections become more organic and you can often see spaces and places where research collaborations then may come. And so, I often think, sometimes serving our academic mission and finding ways of teaching and training that are interdisciplinary then often roll over into our research

mission.

>> DEBBIE CHENG: I actually think that's a great point. And I was recently reading about a health data science degree program that is a joint partnership between statistics and computer science, and it sounded like a really strong and productive program. Anyone else want to jump in with some comments?

All right, if not, I will move next to the topic of data science education. And I have a set of questions on this topic, if that's okay with you, that I'm going to bundle together. So, we've been seeing a sharp rise in degree programs and the number of universities awarding degrees in data science. So, students clearly want more data science, and they want degrees in data science.

There seems to be wide variation, though, in the curriculum of the different educational programs out there. So, the questions I have for you all are: What skills and competencies are essential for the next generation of public health data scientists? And how can educational programs adapt to meet the evolving needs? And then, perhaps more broadly, what do we need to do to ensure high-quality education in public health data science now and going forward? Dr. Hswen, would you like to begin with some of your thoughts on these questions?

>> YULIN HSWEN: That's a challenging question, honestly. I know there's been a lot of talk in the realm of kind of like medical education, especially, too, just because of, as I mentioned, things like ChatGPT being able to pass the medical examinations and so forth, as well as kind of its usage overall in terms of providing kind of clinical advice and whether or not students should be learning to use ChatGPT as a supplement to provide, you know, to provide advice, for instance, to patients.

I guess the first kind of, like, question -- the first thing, I mean, as scientists, that we should be doing, we should be evaluating it, and we should be evaluating its impact on students' cognitive abilities and their ability to provide the best quality of care, and whether or not it's the same type of care. For instance, like in the realm of medical school and to ensure that there isn't what we call automation bias, which is, basically, when you get too used to AI or machine providing you answers that you automatically just assume everything it's saying is correct. But we do know that oftentimes, AI can hallucinate and provide incorrect answers. And what happens in that situation? And I think the issue is that with these kinds of new AI tools, testing them is also really not that reproducible, because they give different answers every time. I guess that's the first answer, is to kind of still try to do those type of evaluations before we deploy them.

In terms of the evolution of, you know, whether or not we can keep up with these types of methods, I think we do have to start using them. I mean, it's not like this type of kind of technology and so forth in this realm hasn't come in. So, there's Wikipedia, there's Google, there's Google Scholar and so forth, and plagiarism and publication bias and all these other things that have come up with the Internet and so forth, and we've adapted to them because we've put in protocols and we've put in kind of like regulations. And I think that's kind of another step forward as well in terms of making sure that, as scientists and educators ourselves, we are using these tools and we are on the pulse of what they're being used for, and we're generating those type of guidelines. So, I hope those are kind of two answers to that kind of broader question.

>> DEBBIE CHENG: Thank you. Thank you, Dr. Hswen, for your comments. I just want to clarify, I really wanted to focus this question a lot more on the educational aspect of public health data science and thinking about, you know, really, what are the essential skills and competencies for the next generation of public health data scientists. So, I'm wondering if anyone else wants to comment really on what we need to do to ensure high-quality education in public health data science.
Dr. Witten, would you like to?

>> DANIELA WITTEN: Yeah, thanks. I think in a lot of ways, this is sort of what we've been doing as biostatisticians for a long time. Within a biostat degree program, you learn like this very classical mathematical statistics training, but you also learn how to collaborate with scientists. Someone with a PhD in biostat will never be an expert on fill-in-the-blank disease or biological area at the level of someone who actually did their formal training in that area, but what a biostatistics PhD can do is learn how to communicate, what questions to ask and learn how to learn. I think that's really what sort of both sides of the aisle need to be doing here, whether the scientist or the statistician. So, I think that, actually, a lot of scientists are not trained in the same way to collaborate with statisticians and computational people in the same way that statisticians are trained to communicate with scientists, and I think that there are some opportunities there for us to, you know, incorporate that training in a typical scientist or public health researcher's training. So, I just want to be clear. I'm not saying that a public health researcher or a biomedical scientist should become an expert on statistics. I don't think that's feasible. I think they've got a lot to learn within their own domain area. But what a scientist or a biomedical researcher or epidemiologist can learn is how to frame questions and have a dialogue with statisticians how to sort of learn from statistical expertise, how to identify cases where they really need that expertise, versus cases where they can do it on their own. And I think that that does require formal training. It's almost like statisticians are formally taught how to engage in consulting with scientists, and maybe it's scientists could also engage in formal training on how to get help from statisticians.

>> DEBBIE CHENG: Yep. And since you brought up biostatistics and our training, maybe I can add a follow-up question. You know, as I mentioned all the interest in data science programs, I'm wondering what you think growth in data science programs, how might that affect how we approach biostatistics education in the future? For example, should they be different degree programs or different tracks, say within a general training program? Dr. Kosorok, do you have comments on that?

>> MICHAEL KOSOROK: Sure. This is a question our department has been grappling with. As the field gets bigger and there's more data science, biostatistics changes pretty rapidly. I think what we're doing now is very different than what we used to do. And we need to be able to pivot, because one of the things about biostatistics is we should always be on the cutting edge of science, right? So, we can't really be complacent and say,

here's a codified set of skill sets we need to have.

And I like the idea of having specialists in different areas within biostatistics, but there's also room for specialists between fields. Like Daniela mentioned this idea of, the importance of having other scientists train on how to collaborate with biostatisticians, but actually, there are some things that are emerging that are really challenging and probably require more than just two people to solve, you know. They may require somebody who actually lives in a space that's partway between one of these biomedical area and statistics who -- for example, a physician whose focus is on their disease area, but also, they have enough training in data science, they can collaborate at a more detailed level with biostatisticians so they can work on more difficult AI questions that may require iterating back and forth in a number of ways. So, I think we need to allow for the expansion of specialties, specialty areas, and encourage those, and I feel that we need to be prepared for the field to always be changing.

And hopefully, as we work with other disciplines, we can share in the ways that we divide up these specialty areas so that they collaborate and work well with other disciplines that have a lot to offer to us and we have a lot to offer to them.

>> DEBBIE CHENG: That's great. Thank you. I'm going to move on now to some audience questions. We have a few questions that have come in around the use of AI. One question is, how can advanced data science techniques, particularly AI, contribute to more effective public health strategies, considering the ethical considerations involved in leveraging large-scale health data, and what potential pitfalls should be carefully navigated in this process? Who would like to jump in first on that question?

>> MICHAEL KOSOROK: I'm willing to say something, if no one else volunteers.

>> DEBBIE CHENG: Sure.

>> MICHAEL KOSOROK: I think that's a very interesting question, and I would like to say that one of the concerns that I have is the idea that just getting big amounts of data will help us do better science, and I feel that it's probably better to focus on the questions that we think are really, really important, and then look at the data that we have and assess whether we need better data. Because we need to design the way we collect both observational and experimental data better. Because there are certain situations where it may appear that we have a lot of valuable data, but let's say we're looking at comparing different treatments, but in one location, they use one treatment, in another location, you get a different treatment. We would never know what would happen if you swap.

Now, there's such a thing as interrupted time series that can sort of help, but there, you need to really be certain that those populations before and after are comparable.

And so, I feel that there's a lot we can do with this data set, but maybe for it to be most useful, we may have to pair that with a question about asking, are there other questions that are more important than these data sets are suited to answer, or can we supplement these data sets to answer the questions we really care about, and if so, how?

>> DEBBIE CHENG: Dr. Witten, did you want to comment?

>> DANIELA WITTEN: Yeah. I thought those were good points from Michael, but I also just want to say, I feel incredibly

concerned by the idea of people saying, well, ChatGPT did well on an exam, therefore, ChatGPT is as knowledgeable as a medical student, or whatever it is. And the reason for that is because I've interacted in my life with enough humans to know that, like, if somebody gets nine out of ten questions right, they probably know a lot about that field, and also, someone who got into med school probably has good common sense and so on, you know what I mean? There's auxiliary information I'm using to inform my belief system for how medical students work. I don't have that auxiliary information about ChatGPT, so you can tell me that ChatGPT got nine out of ten medical questions right, but I don't have, like, a lifetime of lived experience with how ChatGPT responds to slightly different situations versus what happened to be on that exam. And so, I would not want my doctor to be using ChatGPT or anything like that, like, really, ever, in order to be informing any sort of medical decision-making, because I just don't know how ChatGPT would do outside of that very narrow range of conditions for which it was tested.

And I would say, furthermore, the idea that there are these really, really large-scale models that are trained in a completely opaque way on completely mysterious data, and then we don't know what prompts are actually being used in order to generate the results that you're seeing, because there's a bunch of stuff going on between the model and, like, the chat box that you're seeing. We saw last week with the release of Gemini by Google that, you know, there's a lot of room for things to go in unexpected ways. And I think a lot of this excitement around these models, in particular in the context of things that really matter, like public health, I think a lot of that excitement is extremely premature, in my view.

>> DEBBIE CHENG: Great. I'm going to move to the next question. Melody asks, she says, very curious to hear more about how we can better push disparities and equities in our work. Who are the leaders in this that we should follow?
Dr. Goodman, would you like to take this one?

>> MELODY GOODMAN: Yeah, I think there's tons of leaders and people working in this space, and there's people who have done this work for decades. There are several Schools of Public Health who are starting anti-racism centers. We have sort of a consortium around those.

I would also say that, it sort of tags back to one of your other questions -- I think ethics is so important in our field. And so, it's important for all of us to be ethical, regardless of the area that we're working in. And I think I often teach my students that as a biostatistician of data science, you have to be like the ethical beacon on the research team because you're touching the data; you're the one who has access to, you know, make decisions that could potentially -- and sometimes, a lot of times for black and brown people, mean the difference between life and death, right? And so, we want to make sure that our students and any analysts really understand, like, the limitations of the data they are using.

I think someone previously talked about, like, what questions can that data answer? What are appropriate questions to ask using that data? And when are you sort of out of bounds? And I think the technology is allowing people to push bounds, I think in ways that we need to really be careful of, because you have more computing power; you have the ability to do things that you couldn't do before; but I think to Dr. Witten's point, we need some human and compassion and some thought behind this that machines can't give us.

>> DEBBIE CHENG: Very, very well said. Lots of different ethical considerations have been brought up, and I wanted to ask a question to touch on an aspect of ethics that maybe we could talk more about. So, I'm wondering about the ethical considerations with regard to the collection, the use, and the sharing of public health data that we need to address and how can we ensure transparency, privacy, and data security, while at the same time maximizing the utility of the data that we have available? Dr. Hswen, would you like to speak to that, maybe?

>> YULIN HSWEN: I actually am very nervous about all of it, truthfully. I think that it comes from the fact that, kind of speaking, too, about just minority type of populations I think is, one of the issues is that when you're in smaller numbers, I think the likelihood of being identifiable is much more likely. So, I think there's a couple of cases now, I think with -- I'd have to look it up again -- but with like 23andme, for instance, and like the leakage of data and so forth, and with that type of leaking of information, for me on like a personal type of level is very nerve-racking and very scary in that sense, again, because, especially, obviously, vulnerable populations and people with stigmatizing diseases, but also just people who, where the population has like few data points, I think it's very concerning that that type of, like, privacy can, you know, really be broken.

I think the other aspect is, again, is kind of just the overall kind of data usage and the lack of kind of -- I put it in there as kind of like democracy of data is that, you know, we are using data from the public, and that's ourselves, and yet, we rarely speak to the public about how the kind of data, they want their data to be used. And we don't really have them involved. I think that sometimes we do, and we should have them involved in terms of, you know, the research that we're doing. But definitely, these large companies, I'm pretty sure don't do that much of it. And they constantly are using our data all the time. And I think there's a lot of issues with it, and I don't know how comfortable I feel anymore about giving my data anymore in general. So, that's my kind of two cents about it.

>> DEBBIE CHENG: Thank you. Anyone else like to comment? Michael?

>> MICHAEL KOSOROK: Yeah, that's such a complicated question with a lot of different issues in it. I do want to say that I like the idea of thinking about healthy democratization of data science. There's a wonderful group called AI For All. I don't know if any of you have heard of it. It's AI4All. It's a group of concerned young people, high school students. They want to see artificial intelligence used in such a way that it benefits everybody, including all underrepresented groups.

And they think about how this could be done beginning at the high school level. And I really like their vision. So, if you ever hear from them, I think it's worth reaching out, interacting with them, but they will often be paired with a university. So, for example, I first found out about them by working with a high school that partners with Stanford University. And I really like their vision and the way they think about this issue, and I like the idea -- we're talking about scientists now -- how can we make critical thinking, data science critical thinking, part of the conversation for everyone, not just us, so people don't have a huge gap between -- part of the challenge is, you don't necessarily trust what you don't understand. Can we do more to help everybody understand at least the basics? They don't have to understand the deep theoretical underpinnings, necessarily, but they can understand the concepts, and they can be involved in some of the decision making and in some of the research at an earlier stage. So, just a general thought.

>> DEBBIE CHENG: Appreciate that. Turning to the next question. So, during today's presentations, we saw a few examples of how data science, AI, can be used to address public health challenges. There are a growing number of emerging global health challenges, like outbreaks, like pandemics, climate change, antimicrobial resistance and so on. These continue to evolve. And because of that, we need to continue developing and applying innovative approaches within public health data science so we're able to tackle these incredibly complex problems.

So, the question is, in what ways do you see public health data science evolving over the next decade, if we want to be able to more effectively address these different emerging global public health challenges? Dr. Lo, you touched on this during your presentation. I wonder if you wanted to say a little bit more on how you see the next decade evolving?

>> NATHAN LO: Yes, yes. These are all really ambitious and challenging questions. I'll maybe touch on two points. I think the first point is, for any of these problems, whether infectious diseases, antimicrobial resistance, kind of more my wheelhouse, or other issues such as climate change and all the other kind of existential threats. I think in my mind, one of the first things is thinking about the data collection platforms and all of the challenges and considerations in establishing, you know, as robust as possible, data collection platforms that address a lot of the challenges and limitations that have been brought up.

I know in my field of infectious diseases, the standard approach to data collection predisposed us to a lot of issues, including inequity and bias and thinking about from the very start, you know, what is the data collection systems in place for all of these different problems. So, I'd say that that's probably the first point where I would land.

The second point that I'll say is that I'm increasingly interested in thinking about much more rigorous -- or essentially, what is the approach to validation of a lot of models? So, in the perspective of infectious diseases in public health, we've historically thought of validation of models around predictive accuracy in many respects, but I think increasingly, there's a question of, you know, if we use models as decision support tools, should the validation be actually measuring how these models are used and the outcomes from those choices, and increasingly using these models and embedding them within the agencies or individuals and stakeholders whom they're relevant to. And so, I think to kind of realize their potential and impact, recognizing, you know, the many limitations, the considerations that we have to be careful about at every step of the way, I think kind of redefining the way we think of validation is really important.

>> DEBBIE CHENG: Excellent. Excellent points. Thank you. Dr. Witten, would you like to add any comments to that, build on that?

>> DANIELA WITTEN: I think that was a great summary. I think that one thing that I've learned as a statistician during my career is that, you know, we have to respond to the data and the questions as they arise, and we can put together a plan for, like, what we think's going to happen and what we think the challenges are going to be, and the field moves fast enough that we will be wrong. Like, whatever we prognosticate for how things are going to look five years out is going to be in retrospect kind of funny and silly. So, what we have to do is be able to move fast and cope with all these challenges, whether they're LLMs or, like, disparity in data, or whatever it is, as they come. So, I'm excited to see all these people here on this, as part of this conversation, who want to be a part of this.

>> DEBBIE CHENG: Thank you. This next question is from our audience member, Anthony. He says, I'm from the New York State Department of Health and the Office of Aging and Long-term Care. I was wondering what the panel thought that government agencies could do to become effective stewards of modern data science. We're undergoing a current data transformation using cloud platforms, modern coding platforms, visualization platforms, to both provide live interactable data to key decision-makers as well as begin to be able to truly study the long-term care system that's been underrepresented in the literature.
Dr. Goodman, would you like to jump in first on this question?

>> MELODY GOODMAN: Sure. Seems like they're doing the main things I was saying, making data accessible to those who need it, so the cloud-based things. The second thing I would say is actually speak to users. Often, people putting data out for use don't understand how people actually use it in practice, and that may dictate how you want to display things or making things easier for others to actually use.

And then, the other thing I would say as a government agency is, yes, it's great to make data available for people like us, but it's even better if you can make your data and tools available to lay audiences in ways that they can understand. And I think this is where some of the technology can help. Like, using technology to develop their own visualizations and their own -- pulling their own data for their own communities is a great way government can really help be good stewards of data. Because statisticians, we can figure out how to access even crappy data, but the general public really needs it to be in user-friendly forms.

>> DEBBIE CHENG: Thanks. Anyone else want to add to that? Okay, next question. Michael, you talked in your presentation about the importance of us engaging and leading more from the front. And recently, I had some clinical investigators say to me, I'm still not exactly sure what data science is and how to incorporate it into my work. And comments like this have made me wonder what we should be doing as data scientists to lead more from the front in engaging clinicians and domain area experts to help them understand how data science can advance their work. So, I'm wondering if you all could share your perspectives on that? Dr. Hswen, do you want to begin?

>> YULIN HSWEN: I'm sorry, can you repeat the question again? I'm so sorry. I got cut off.

>> DEBBIE CHENG: Oh, no problem. I was asking what we can do as data scientists to lead more from the front in engaging clinicians and domain area experts who may be unsure how data science can help them advance their work.

>> YULIN HSWEN: Oh, goodness. This is being recorded. I could say a lot of things. I would say -- I mean, in all honesty, I think it's -- I mean, I think it's communication. You know, I think, I know Dr. Witten had said, you know, collaboration, and kind of the importance of it. You're kind of taught to do that. I think we're always kind of coming to them all the time. At the same time, I think that, you know, I think there's more knowledge around the power of data science and the data itself. And so, we're seeing now more and more kind of clinicians and other researchers in other areas kind of come to us for that kind of information, and I think that's just kind of being open to it.

I think it's collaborating in these domains and making sure that you are on projects with a multitude of interdisciplinary researchers and people. I think that's kind of where to start. And trying to kind of communicate best forward as to understand what their question is, but again, for them to kind of understand -- I think they need to understand, you know, what you're doing, and also the limitations, I think, of the data and the questions that can be answered from the data. Because I think, oftentimes, in general, people want things very simplified in like a media headline, and then the rest gets lost, and that's where issues can arise.

>> DEBBIE CHENG: Great. Thank you. All right. Well, we are getting close to time, and I have one last question that I'd love to hear from all of you on. So, perhaps I'll pose that to you now. I'd like to know what your final take-home message is for our audience today, a last take-home point that you'd like the attendees to come away with on what we should do to prepare and help shape the next decade of public health data science.

>> MELODY GOODMAN: I can go first.

>> DEBBIE CHENG: Thank you.

>> MELODY GOODMAN: So, I would say that data scientists are narrators, and it's important to have a diverse set of narrators to tell the diverse set of stories in public health.

>> DEBBIE CHENG: Dr. Lo?

>> NATHAN LO: I'd be happy to go next. I think my kind of summary would be that I think, you know, figuring out and developing a framework for really critical evaluation of your entire process for a question, from the data that's being gathered to the model development validation use across many dimensions, but including equity impact, rigor, and how your information will be used, and developing that framework for yourself and being critical across all dimensions.

>> DEBBIE CHENG: Dr. Witten?

>> DANIELA WITTEN: Yeah, I would say building teams that have the right composition of statistical and data science expertise alongside domain expertise, in order to be able to answer the right questions correctly.

>> YULIN HSWEN: I guess I'll go next. I think it's just don't lose touch with, like, humanity. So, even though we work with data, I think that you still need to talk to people; you still need to work with people and you still need to interact with the public themselves to actually truly know where the data

is coming from and what it actually means. I think sometimes we get lost in just looking at numbers, but we have to remember that it's people.

>> DEBBIE CHENG: Well said. Dr. Kosorok?

>> MICHAEL KOSOROK: (Muted).

>> DEBBIE CHENG: Michael, you're muted, in case --

>> MICHAEL KOSOROK: Thank you. I said some nice things there. I forgot. I think this is a time we have to be really prepared for rapid changes in the scientific field and in the questions. Every point that's made is really important, but I think we have to be prepared to tighten our seatbelts and watch things move and be able to pivot. Learn how to stay up with those things, learn how to not be too pulled aside by overhype of things, as we've been talking about, and then learn how to communicate those things and bring others along. I think that's going to be one of our challenges.

>> DEBBIE CHENG: Great. On that note, I see we are at time now, so I will wrap things up at this point. I'd like to thank each of our panelists, as well as the audience today. Thank you all very, very much for a truly insightful discussion. It was a real honor for me to serve as moderator. I will now turn it back over to Dean Galea.

>> SANDRO GALEA: Thank you, Professor Cheng. And really, I'm just echoing everybody's thank you. Ahead of these conversations, I'm always looking forward to learning from them. I must admit, I learned so much more here than I thought I would learn. Any conversation in data science that is so clear about the role of bias, the role of data science in helping serve our larger purpose, about the role of our humanity understanding data and how data science is all about telling a narrative, is a conversation that's worth listening to and relistening to, and I'm deeply grateful to you all for making those points. Thank you for the work that you do. And I want to thank the audience for a really, really interesting set of questions and conversations in the chat. Everybody, it's a privilege to have been with you for the past hour and a half. I hope everybody has a good afternoon, a good evening, or a good morning, depending where you are in the world. Everybody take good care.

>> DEBBIE CHENG: Thank you.

(Session concluded at 2:31 p.m. ET.)

This text is being provided in a realtime format. Communication Access Realtime Translation (CART) or captioning are provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings.