MILDRED CHO: Good morning! Or afternoon or evening, depending on which part of the world you are Zooming in from today. I'm Mildred Cho. I'm delighted to welcome you to our ELSI Friday ‑‑ (freezes) This forum is hosted by the Center for ELSI Resources and Analysis and held on the second Friday of every month at noon Eastern Time in the U.S. We also have an informal Zoom room for discussion immediately after the panel for 30 minutes and hope you will join us for that.

For those of you who might be new to the center, we provide resources to support research on the ethical, legal, and social implications of, ah, of genetics and genomics. And it serves to connect scholars, scientists, policy‑makers, journalists, members of the public, and others to engage ELSI issues. It's funded by the National Human Genome Research Institute at NIH and is managed by teams at Stanford and Columbia Universities in partnership with the Hastings Center and Harvard University. I encourage you to visit CERA's online platform, ELSIhub.org, for the recording and transcript of this forum and related references.

And I'm pleased to announce the new ELSIhub Collection ‑‑ (cutting up) This reading list explores shortcomings in the current regulatory framework for AI‑based medical devices and key legal and ethical questions about bias, privacy, and liability. Please also go to the website to join the ELSI Scholar Directory, sign up for newsletters and other events like this one at ELSIhub.org, and get daily updates and news on Twitter.

So just some quick housekeeping information. If you want to use closed captioning, please turn on CC at the bottom of your screen. And, ah, use the Q&A button for questions and, um... which you'll find at the bottom of your screen. And you can use that to write that in, and also use the upvote button for Q&A. So again, please use the Q&A instead of the chat if you want to have questions. But you can do that any time throughout the presentations or discussion.

We'll post links in the chatbox for resources referenced in today's discussion. If you have any e‑mail ‑‑ questions, you can e‑mail info@eLSIhub.org at any time.

So, without further ado, I will introduce our moderator, Dr. Lucila Ohno‑Machado, who is an associate dean for informatics and technology and also the chair of the Department of Biomedical Informatics, and a biomedical engineer. She is a member of the American Institute for Medical and Biological Engineering, the American Society for Clinical Investigation, and the National Academy of Medicine. Her research focuses on privacy preserving distributed analytics for health care and biomedical sciences. She's received numerous awards in biomedical informatics. So I'm very pleased to have her here today to moderate our session. Please take it away, Lucila.

LUCILA OHNO-MACHADO: Thanks, Mildred. Thank you for inviting me to this forum. It's very exciting. We were talking about balancing data privacy and data sharing, something that we used to talk theoretically about and now is very much in practice. Very exciting times. AI and predictive analytics require large amounts of data to build and analyze, and the whole world of evidence‑based medicine relies on large collections of data often obtained from electronic health records ‑‑ that is, clinical records collected for clinical visits, hospitalizations, emergency department encounters, and so on. What's interesting: Not all patients know that their records may be used for research without their explicit consent. Not all patients know about the potential benefits of having their records used for research. The new discoveries, the foundation or removal of predictive models that have clinical implications, they are all depending on this data‑sharing. Not all patients, and then not all researchers, know about the risks of using health information, regard to patient and even institutional privacy.

So we have two experts, international experts, in this subject, to have presentations and then a debate after that. And the two panelists are Dr. Brad Malin, who is the Accenture professor of Biomedical Informatics, Biostatistics, and Computer Science at Vanderbilt University, vice chair for Research Affairs in the Department of Biomedical Informatics. His research is to construct AI learning applications in the context of real‑world political architectures. He codirects the Health Data Science Center, the Center for Genetic Privacy and Identity in Community Settings, and also the Infrastructure Corp of the NIH AI Learning Consortium to Advance Health Equity and Research Diversity, the AIM‑AHEAD Consortium. In addition, he's currently cochair of the Committee on Access, Privacy, and Security for the All of Us research program.

Also presenting today is Dr. Effy Vayena, who's professor of bioethics and the deputy head of the Institute for Translational Medicine of the Swiss Federal Institute of Technology, ETH‑Z, renowned expert at the intersection of medicine, data, and ethics. Her health advocacy policy lab focuses on ethical and policy challenges in precision medicine and digital health. She's a member of the Swiss Academy of Medical Sciences and chairs the Ethical, Legal, and Societal Implications research group of a national infrastructure and research program aiming to advanced personalized health in Switzerland. She has previously worked with the World Health Organization, the Welcome Trust, and multiple other organizations and institutions around the world.

So please welcome our two panelists. We will start with Dr. Malin.

BRADLEY MALIN: Thanks, Lucila. I didn't know that Effy and I were going to be in a debate! So this will, ah, this will be interesting. Let me go ahead and share my screen.

Okay. So, um, I was asked to speak about some of the technological aspects of data‑sharing, as well as the benefits and potential challenges associated with their application. So, I'm going to start... by talking about the first technical approach, which I have no doubt is on everybody's mind, which is the principle of de‑identification, or the notion of anonymization in the context of the EU.

So lay the groundwork so we're all on the same page, what HIPAA has to say, is this is information that does not identify an individual and has no reasonable basis for which that information could be used TO identify an individual. And then this gets realized in several different implementations under the regulation, one of which is what's affectionately referred to as safe harbor, where it says you need to remove 18 types of identifiers, and then attest to the fact that you have no actual knowledge that the residual information can identify somebody.

The alternative to this is what's referred to as an expert determination, in which somebody takes statistical or scientific principles and applies this to the data to show that there's a very small risk that an anticipated recipient could identify an individual.

I think that you'll see similar language, if you look at GDPR, in which there's, ah... this notion of anonymization, which talks about the relationship of the data to an identified natural person.

Okay. So how does that actually translate into practice? So, one example in which we're doing this is within the context of the All of Us program, which I believe many people in the audience are familiar with, but. For those that aren't, the All of Us program began several years ago and is working to create a national representative cohort of around a million individuals who have contributed medical records, data, bio‑specimens, as well as participant‑reported outcomes.

And what we've done within All of Us is create a tiered level of access, of which de‑identification is sitting at the heart. First, though, the first tier is we've created a public access resource. And this is information that can be accessed without logging into the system ‑‑ mainly summary statistics about the types of participants we have, in terms of their, ah... race, age, sex, gender distribution, as well as their diagnoses and, ah, medications and labs and anything that you could imagine would show up in their records. This is all summarized information. We then create sandbox environments, which are sitting out in the Google Cloud Platform, or GCP. The first tier here is what we've referred to as registered, and this is accessible to anyone within a trusted organization that enter into a relationship with, with the program, so that there's some type of a assigning authority that accepts responsibility for the organization's involvement. Um. And then there's a user code of conduct that individual users enter into with the program to, to gain access to data.

And as I'll show in a second, this is data that we deem to have low risk of re‑identification after a statistical assessment.

Last week ‑‑ maybe a week and a half ago ‑‑ we released what is now known as the controlled tier. And this is information that, again, at the present moment in time, it's available to researchers in trusted organizations, but we've incorporated the genomic data, whole genome sequencing, on a large portion of our population. Ah, and this also has more detail in terms of some of the demographics. I, I would like to caveat this by saying that both of these environments are designated as nonhuman subjects environments. And thus have a de‑identification exemption.

So, a lot of people ask: So, what is actually in this resource now? And what have you done to the data in order to de‑identify it? And so, this is, this is a quick summarized view ‑‑ 'cause I don't want to spend too much time on this ‑‑ about what you'll find within a registered tier, versus what you'll find in the controlled tier. And so you can see that some of the differences is that in the registered tier, which is what we're expecting to be a more widely disseminated environment, is that we've taken the dates of events and we've shifted them randomly over the course of a year. We've generalized the geolocation in terms of where an individual says they reside, well as certain aspects of race and ethnicity to obscure low counts for populations that we just don't have a sufficient number of participants about yet. And you can see there's a lot of other information that we've chosen to suppress! And so this is really at the heart of the notion of de‑identification, in that in order to achieve sufficient protection, there is, there's information about an individual that we cannot disclose.

Controlled tier, there's more protection, ah, in terms of the controls for who's allowed to come in and work with the resource, and that's more detail that gets put in here. And that's where I alluded to the genomic data has recently been released.

The same type of premise has, has actually been used for other ‑‑ many other programs, both within the U.S. as well as outside of the United States. And I wanted to draw everybody's attention to what's been going on in Europe, with respect to the European Medicines Agency as well. So under policy 70, they indicate that you can publicly disseminate clinical trials data... ah, and then in the guidance that they issued about five years ago, they drew a number in the sand and said that the risk of re‑identification for individuals within these trials can be no worse than 0.09. And so I would encourage everybody to take a look at this, this documentation, if you want to know more about it.

Other examples in which this has occurred: Project Data Sphere is a problem that was established almost ten years ago to support the publication of various clinical trials, potentially ‑‑ particularly in the oncology space. And, and so there have been dream challenges that have been built around this resource. This has been a resource that has, um... really been focused more on failed phased re‑clinical trials data. But it's supported the reuse of this data in innovative, novel ways. One example of which is taking a look at the natural etiology of prostate cancer across multiple trials on a relatively large scale that had not really been seen before.

So, so there's opportunities in this regard, in terms of de‑identification. But one of the drawbacks to this is that you do end up hiding information from, from the end user. And so there's, there are questions about whether or not the types of research that we want to support moving forward, in terms of disparities investigations or using really detailed social determinants of health, fit within to the confines of this type of a technology. There's also some concerns that as the amount of information that we start collecting and sharing grows, that it becomes increasingly more challenging to, to offer guarantees of protection.

So, more recently, the, an alternative to doing de‑identification has been suggested in the form of what's known as secure multiparty computation, wherein in this situation you don't necessarily disclose information in the clear to the end user. Instead, you would take a model that somebody's interested in using for, for learning off of the data, and then encrypting that and sending that over to another party, where they can then encrypt their data, and then you can run an encrypted analysis between the data and the model that is going to be applied, and learn aggregate results and summary ‑‑ or, not summary statistics, but the models associated with the predictions themselves, as well as the confidence intervals with the predictive models. This has been used in a, in several different environments? And, ah, we can go through them if we have time later on.

The most recent, I would say, is what's called a secure enclave framework, where in this situation everything gets moved not into a encryption itself, but into a secured environment on a computer in a tamper‑resistant space. So it's encrypted in the sense that, looking at what's going on within the computer, you can't tell what's going on. But there is a secured ship, secured space in terms of secured RAM. Specifically examples of this are Intel's SGX or software environment that will allow you to take ANY data, encrypt it, push it into the SGX, decrypt it, and then process it in an environment that you don't have the ability to see what's going on, but then you can encrypt it and take the results back out at the end. What's notable about SGX is that this is now embedded in just about every computer that, that we come across today. And that Microsoft and, ah, Amazon are now supporting server farms for processing data out in their, um, their public compute environments.

The, ah, one other technology that I wanted to allude to was the notion of synthetic data. And this has been brought up a lot, over the last couple years, because, ah, mainly because of the notion of the concept of deep fakes. Where, where it's pretty obvious that with respect to imaging that, ah, you can... create, um, realistic views of an individual. And so the potential application of this in the medical space is not really new! We've been doing this for, for a while now. So we had a paper that came out about five years ago talking about the, ah, creation of synthetic clinical data. Which, we applied this to a combination of demographics, diagnoses, procedures, and meds from data from Sutter Health and MIMIC. And then just to give you some intuition into, like, does this work, what you're looking at here is the rate at which diagnoses or these clinical concepts show up in the real data and the rate at which they show up in synthetic data. And the expectation is that the synthetic data, because it's not, it's not corresponding to any particular individual, that this is sufficiently privacy‑preserving. But there are, there are some questions about whether or not, ah, it's applicable in practice.

And, so. Just as an example of things that you could do with this ‑‑ and I'm glad that one of my former grad students and current postdocs, Chai Yen, is in the room, 'cause he assisted with this. We developed versions of the All of Us data that could be used in demonstrations and tutorials, synthetic versions. And just a demonstration of what you're looking at here is on the right, you're looking at some of the height distribution of the participants within the program that you see in the real data. And then. So, for context, EHR is what comes out of the medical records; PPI is the participant‑contributed information, or participant‑provided information. And then on the right, what you see is the synthetic data. And they are approximately the same. This is not always going to be the case, but it's an illustration of how you can provide data that has the look and feel of the original data, even though it's synthetic.

That said, there are concerns about what synthetic data actually allows for, because there have been investigations to suggest that you may be able to rewind or go back to the original information, because the synthetic data is going to be not exactly the same as the individuals whose data was sent in for training the model that generates the synthetic data. But you can see that, in this situation, you end up with things that look fake, but they're sufficiently similar to the individuals that they came from such that they're, with some prior knowledge about who COULD have been in the underlying training dataset, you may be able to invert the, ah, the program.

And so, so, if a user can test if someone knows that somebody is in the dataset, then they may be able to do such a rewinding. Now, the challenge here is that they would require knowing certain features about the person. And also, they don't necessarily learn anything new! The learning, the violation of privacy at this point would really happen with when you can show that somebody's IN the dataset, but you can also infer something NEW off of them based on the information that's been published in this dataset. As I alluded to.

Um, one final piece, and then I'll yield the floor, is that this notion of, coming back to this notion of secure computation... there was this ‑‑ I was asked this question recently about like, you know, can this be put into real working software? And the answer is, ah, it can totally be done, and it can be done in a relatively efficient manner. This is just an illustration of work that came out of Jean Louis Raisaro's group, which looked at how you can share COVID models and COVID data in a secure fashion across international borders. And so I would encourage people to take a look at this if they're interested.

So, so in closing, there are still things to worry about here. So the quantity and diversity of the organizations and the users is continuing to grow, which make things challenging for protection users. As well as the fact that the data is increasingly more detailed and complex. Which begs the question of if the models that we've developed to date are really gonna be applicable to the models that we're going to need moving into the future. And this notion of risk assessments, and saying whether or not data is sufficiently de‑identified or sufficiently protected... they help, but they tend to make assumptions about what's an acceptable level of protection within society. And coming up with agreements about this ‑‑ I gave you the allusion to the 0.09 from EMA ‑‑ there IS no actual number that somebody says is, this is a sufficient level of protection in order to achieve compliance with HIPAA or the common rule in the U.S.

And finally, I want to note that breaches will happen! And, and that's something that I believe that we're gonna need to, ah, accept? And that there will need to be responses that we have prepared. And we have to have an understanding of what is, ah... a magnitude of a, of breach to the point of which, you know, we claim that the organization was at fault, as opposed to which there was an incidental disclosure. Regardless, I believe transparency is gonna be critical for this.

And with that, I will yield the floor. I realize I went a little over time. But I will allow Effy to, ah, rebut what I've said.

EFFY VAYENA: Thank you ‑‑ thank you, Brad. And I'm not intending to do that! Because you and I agree on so many things, it's hard to debate! I'm just gonna take a different, um ‑‑ I'm gonna talk about this issue from perhaps just a different point of view. So, just let me... I think I can share... Could you let me share, I think ‑‑ yeah. There we go, okay.

Okay, so I need to share my screen, and. Right. Hopefully that works. Here we go.

So. Yes, again, so it's not ‑‑ I'm not going to argue with you! But I'm going to argue that what we've been trying to do is what I call this quest for health data governance, which is the, another way to look at privacy.

So, I'm, I've used recently this report from the OECD, in the series of the Health Working Papers, and I'm going to refer to this a little bit just to illustrate what I'm thinking about about this matter. So of course, the usual points about how important all the data uses are and the transformation that it's gonna come for our health sector through the uses of data. But if you look at this list and the bullet points of what is going to happen, you can see that our expectation is to do so many different things. Our expectation is to respond to public health challenges, to improve health care systems' performance and health care quality, to do scientific discoveries, and to save lives. And on the graph on the right, you can see that what else we're expecting from this whole thing is to have more cost‑effective, um, health care systems. It's not just about their quality and their performance, but also it's gonna save us money, and it's gonna save us lives, again, because of the... moving from treating to prevent and predict.

So, all that means is we're actually talking, always, about different kinds of data. We are talking about multiple purposes of uses here. And we have many stakeholders involved. So the system is already very complex. And Brad already alluded to that increasing complexity. And that, for me, is an important point to remember when we're trying to balance things such as privacy and access, because these different kinds of data, these different kinds of purposes, we will have inevitably, I think, different thresholds for that kind of balance.

So, in trying to find that balance and develop frameworks and, and ways in which we can achieve our purpose, we've been on it as a community, and our ELSI community as well, for a long time. And this is some work that we did some years ago already looking at what has happened, how do we theorize, or how do we actually p‑ ‑‑ what kind of practical frameworks have we put out there, in order to do this balancing between the need to protect privacy and to be able to use those kinds of datasets we have? And what we found in our report, in our work, was that in those many, many documents that we examined ‑‑ first of all, there's been a number of frameworks, which I call outputs. And I'll come back to the outcomes in a minute. What we saw is that what we've tried to think about in these frameworks is mostly privacy, is mostly autonomy, and that's primarily in the form of informed consent that gives the opportunity to people to make decisions about what they want to happen with their data. And another one that came out was that we were concerned in this process of thinking, developing the frameworks, about data quality and curation. So these are several outputs.

But when we look at what happened in 2021, we are still having the OECD telling us that in order to get to this transformation and use this, what we are still in need of is effective health data governance that enables secure and privacy‑protective data uses. So it seems that despite all the effort that's been put in there, we haven't maybe reached that point of effective health data governance. But also, that we want to do that by paying our maximum attention to privacy and privacy‑protective data uses. So the ‑‑ it seems to me, and that's the point I'm trying to, will try to make: That in ‑‑ we have our frameworks and our thinking about finding a balance, is... having privacy as a central, a very central piece of it.

In the meantime, I'd like to introduce to you that interesting p‑ ‑‑ two interesting pieces of work. One is the paper by Ibrahim et al which introduces this concept of health data poverty, as the inability for some groups or populations to benefit from the discoveries that they could benefit from, because ‑‑ due to the scarcity of data. And, ah... the usual problem, that data is not representative. And keeping that in mind, the paper is in 2021, so we're still suffering from health data poverty, despite all the outputs that we've put out there in terms of frameworks and ways in which we want to solve this problem. At the same time, this year, early this year, we had the Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world, that actually made the further claim that if we want to improve equity and health and wellbeing, those digital technologies and digital transformation through increased access to data is gonna play a crucial role. They went even further to call the digital as a, um, actually a determinant of health, adding to the known social determinants of health.

So, the outcome from all the effort that's been put on the frameworks for using data to improve health equity still gives us a problem that we call health data poverty.

AND, even in countries where we went ahead with, for example, regulation that ought to be up‑to‑date ‑‑ landmark regulation, like the GDPR; I'm sure people in this room here are familiar with that ‑‑ we've still been struggling with some of the problems that we initially wanted to address through these regulatory tools. A paper that you will see ‑‑ you've probably seen those papers before. GDPR has created frustration in global biomedical research. Again, with issues based on privacy. But also what we saw recently during the pandemic was that we had problems, in terms of sharing and access of data, that were not necessarily only a privacy problem. But we had other kinds of problems that frameworks that have been very focused on finding the balance between access and privacy haven't necessarily addressed efficiently or effectively.

To the extent that we had ‑‑ I find this quite a striking thing to read in 2021, in the middle of a terrible pandemic: that we still did not have rules for pandemic data access that everyone can trust. And again, I don't think here we're speaking about data that, necessarily, and only about personal data where the privacy questions that we're worried about apply.

So, we have ‑‑ I'd like for you to keep in mind that we have this issue with outcomes, in terms of frameworks, technological solutions. But the output is still an issue.

And I wanted to bring you back again to that paper from the OECD that looked at what countries have been doing in addressing this problem, not only in terms of theory by developing a framework, but actually in terms of application. Getting it to work in practice. This is, these are data from 1920, from a number of OECD countries. You can see that they are looking at the national health datasets availability, maturity, and use. So these are perhaps some arbitrary criteria and indicators. But here, you can see what they are on the right. It's about their actual, the availability of the dataset; what they cover; whether there's automation in some of the retrievals; the timeliness; whether there's unique identification; data linkage; et cetera. Again, I recommend to take a look at this quite, I think, interesting to see.

Now, you can see here that we have countries like Denmark that seem to be doing very well, according to these indicators. Countries like the United States, NOT as well, according to these indicators. So, is our conclusion here then that looking at ‑‑ in trying to implement those frameworks still, we have the problems despite, again, looking ‑‑ despite the regulatory forms AND the technological developments in the meantime?

This is an, also an interesting slide. If you look ‑‑ I can't go through it now, in the interest of time, and I'm hoping you can see a little bit, it's big enough for you to see. The, what it shows is the kind of governance difficulties that we have. And again, what I want to point out is we have a series of legal and policy barriers, but we also have some other things that we, maybe we don't pay so much attention to. For example, the personal, the person identifiers that links the data. The quality of data. Technical capacity. And a bunch of other things that seem to be a problem.

So... I'm gonna use an example that we ‑‑ so, we have the problem that we haven't been able to find a good balance between privacy and use of data, or that I think we need to rephrase the problem ‑‑ and that's what I'm trying to say today ‑‑ that perhaps what we don't have is not necessarily that balance between privacy and access, but we don't have these appropriate health data governance, which is a bit bigger than just the privacy versus access. In Europe, we are trying now a new thing, and we are ‑‑ the EU has put forward what we call the Data Act. The Data Act ‑‑ again, the GDPR's still in place and applies. But the Data Act is trying to clarify who can access and share data, and on what terms. And the Data Act, it will now attempt to provide legal certainty, and the aim is to remove barriers to data sharing. For those of you who've been around for a while, you've heard that rhetoric for, for many years; we all had. But now, at least in Europe, the idea that we... IS, that we can achieve that kind of clarity through regulation, but also through this thing that we call European Health Data Space. And the idea here is that we're gonna put together somehow all different kinds of data. It's a project the EU Commission started already in 2019. No data have input anywhere; it's still thinking and processing, but that's the timeline. And it's going to include all kinds of data with, like ‑‑ again, we've been thinking about for so long electronic health records, genomics, patient registries, et cetera. The idea that is this international, within Europe, and the regulatory tools and the regulatory instruments that we have in Europe will apply. But the aim will be to use the data, versus to, you know, block the data use.

And here's some interesting things from the plan, at least. Again continuing with the hope to have better health data governance. First of all, what we define as health data. You remember in the beginning, my ‑‑ I started off by suggesting that we have WAY too many purposes, WAY too many... to diversity in the kinds of data we're using, and all of those, they might have different risks and different, ah ‑‑ therefore, the thresholds will be different. We, within this scheme of European health data spaces, we are going to try and solve the interoperability problem, especially for cross‑border exchange. The security will be a priority. There's gonna be focus on secondary uses and scientific research. And here comes something which is quite interesting: The suggestion is to introduce also something that's called data altruism. And that actually means a consent, a type of consent... whereby the person will be giving authorization for altruistic purposes, without, you know, expecting something back, and it will be perhaps something closer to what we have called "broad consent" in the older days. It will also involve national authorities having the oversight of what's happening in its country in order for the data to become part of this European Health Data Space.

So, I think what we're trying ‑‑ what happened recently in Europe, which you may or may not have followed, is we had this extreme enthusiasm during the pandemic for digital contact tracing apps. And we... most of the European countries actually introduced one. It was run on Google API. But what was interesting about this was the way we debated privacy. This is, of course, given the GDPR, given our challenges with privacy versus access ‑‑ even within a context of a public health emergency, we've been perceived ‑‑ and this is an interesting paper that actually describes the Europeans ‑‑ as those who were not amenable to sharing location data for contacting tracing in health emergencies. But the last line of this highlighted text, which is my highlight, is that it should help ‑‑ the authors say here, it should help in combating the next pandemic that the balance between preserving privacy and preserving life has changed during this pandemic. And I think that, for me, is the epitome of how perhaps wrongly we frame this question. Because I don't think that in Europe, we went for preserving privacy instead of preserving life. It's this kind of framing that everything around data access has something to do with privacy. In fact, in Europe, we did use a very privacy‑preserving approach to this particular application. We had a fight which resolved it, and it was privacy‑preserving at the end. But then, even with the privacy‑preserving approach, which was quite effective ‑‑ if we had all used those technologies ‑‑ we saw that in Europe, the ‑‑ and you can see here the penetration and the percentage of the population that used those apps ‑‑ was actually not spectacular. We have only one outlier, which is Finland. But everyone else was kind of lukewarm about those applications. And the point is, that I'm trying to make, is that it wasn't because these systems were not privacy‑preserving. We got that okay, almost quite right. Something else was the issue.

And so, in, in closing ‑‑ not eating more time of our discussion ‑‑ I think... the question of data governance, despite all the effort that we've put in, is still a work in progress. But I think it's being framed in a way that hasn't been as constructive and perhaps as helpful in actually finding that balance! We frame things as privacy versus, you know, everything else that's wonderful. And I think that's... problematic. There's so much more in here that we need to pay attention to in order to get it right. And for that, of course, we have the technical part ‑‑ and Brad told us about how many different pieces of technology we can use today to get that part right ‑‑ but also, the governance mix that is, the governance part that I think stretches beyond just the privacy pillar is what is required. And although we've been, again, talking about these other pieces that will complement governance for, again, some time, I think now it's time to get on with implementation instead of adding more theory into what the governance, ah, good governance might be.

So, I'll stop here, and I'll stop sharing. And I'll yield the floor back to Lucila. Thank you.

LUCILA OHNO-MACHADO: Thank you so much. Very interesting presentations. And I must apologize for the choice of words. Not being a native speaker, "debate" for me did not mean opposing arguments. I did check on Oxford Dictionary afterwards. So, I actually meant discussion. So, thank you so much. And I think the audience will agree, these are not opposing views; they are actually very balanced views of this, ah... balance between making the data useful and ascertaining privacy of the individuals.

So we have already two questions on the box here that I think we could get started with. I just had a couple of clarifications. One is for Brad and the magic number 0.09. Ah, is that 9%, or is that 0.09%? What is that?

BRADLEY MALIN: Yeah. There's a long history on this. I'm, I'm not going to go through all of it, but. So the short version of the story was that this number initially was derived from the way that CMS provided guidance on how data could be publicly disseminated. And their guidance was that they, when sharing data that you might get from like a multidimensional contingency table, or if you just wanted to say like a cell count, they said you can't publish anything that has a count smaller than 11. And, so then there was this discussion around, well, what does this really mean? Like, do we want just cell counts, or should we equate this to risk? And when the EMA was saying, well, we want to use risk as the definition, because this was more in line with the way that GDPR had defined privacy protections or the notion of anonymity, then they translated this into 1 over 11, to get risk of 0.09. And so, that's what it means.

Now... can you translate this into something meaningful, in terms of like... (computer chime) Does this mean, for instance, that if you put everybody in a group of 11, that we'd be willing to tolerate that if somebody randomly guessed the corresponding identity of those individuals, that they should be allowed to be right 9% of the time? I'm not sure that's what the intention was! But it seems as though that that seems to be a potential implication of it.

LUCILA OHNO-MACHADO: Mm.

BRADLEY MALIN: I do think, though, that their intention was that for any record independently, that the risk of de‑anonymizing them should be no greater than 0.09. Given all of the protections that get put in place, whether they be data removal, or whether it be some additional deterrents, such as some type of a controlled access environment.

So I believe that that's what they're trying to say, with that number.

LUCILA OHNO-MACHADO: And... it's very ad hoc, yeah. (computer chime) That's...

BRADLEY MALIN: Well, the number itself! I mean, this is the problem in general. No matter what ‑‑ there is no magic number! Right? And so different agencies around the U.S. and around the world choose different numbers. So there are some states, like some public health reporting agencies, they choose a number of 5. There are other ones that choose a number of 20. CMS chose a number of 11. You know, like who's to say what's the correct number? I think what's more important is that it's not 1. Right? (chuckles) You know, it's, um! There is some level of protection, but. I don't think that there's really ever been a universal agreement of what is going to be the right number. So. I don't even know how to get to that point! It's really a fascinating opportunity, for discussion. But, um... you know, it's ‑‑ one piece ‑‑ I really love how Effy focused on this notion of governance. Because it is just one piece of the governance aspect!

LUCILA OHNO-MACHADO: Right.

BRADLEY MALIN: Which is that, well, if you incorporate a technical component INTO the governance, how do we come to an agreement of what is the acceptable level of risk with that technology?

LUCILA OHNO-MACHADO: Mm‑hmm.

BRADLEY MALIN: So that's, that's there.

LUCILA OHNO-MACHADO: Yeah. And even though the indicators, they are equally weighted in... (chuckles) Just to rank the countries, ah, as it was being portrayed. And probably, there is a corresponding graphic, graph, about the risk. Right? Because that's the balance we're talking about.

So why don't I go in kind of distribute the questions to both of you. So, maybe start with Effy answering Barbara Koenig's question: Why is it so difficult to shift the focus away from an autonomy framework focused on privacy and informed consent, and toward careful governance? Effy.

EFFY VAYENA: Thank you. Thank you, Barbara. It is a good question. We've been struggling to understand. I have, perhaps, a... I don't know, I think it's more of a psychological, I guess, analysis of why we're doing that! But I think we are probably... ah, culturally thinking that it's one piece we can, you know, regulate. We can fix, somehow, a rule about privacy. We can have a form about consent. And I think it's ‑‑ we simplify things, and we're stuck in that... almost redactionism, I would say! To get where we want, we want another piece. So it's easier to think we'll solve our problem with one solution. Which is, okay, let's have privacy, let's fix consent. And it will all magically work out. And we know it's not gonna work.

So I think we kind of ‑‑ there's an inertia, that having thought that this one piece will solve the problem, we're not opening up to understanding the rest of it. And what struck me in the work we did, it's not that we haven't thought about it! People like you and so many others have been writing about this issue for so long! But we're making, we have made, I would say, incremental progress! So I think we're held back by those notions of, not just not wanting, but then we use it also a little bit as an excuse? But the simplicity that comes ‑‑ or simplification or redaction that comes in terms of, we have one problem to solve, instead of all the others. I think that's probably the explanation.

So I'm very curious to see what's going to happen with this new piece of regulation in Europe, and the effort to create those European shared data spaces, the European Health Data Space. Because there, the idea is to move away from these, you know, uber focus on privacy and move to the other pieces. Throwing in consent, but it's more than that? So I'm very curious to see if that can move the middle. But, I don't know. It's an experiment I guess.

LUCILA OHNO-MACHADO: Thank you. The next ‑‑

BRADLEY MALIN: Mind if I chime in on that for a second?

LUCILA OHNO-MACHADO: Sure.

BRADLEY MALIN: Just ‑‑ I think there's a really important part to this equation, in that... you run into this problem that we're trying to conflate all different uses? Into one scenario. And it's, and I think this is what Effy was getting at, which is that it's NOT one scenario. There's a big difference between supporting public health research, versus sharing clinical trials data. And trying to create a one‑size‑fits‑all policy model for this is... it leads to a lot of frustration and confusion, because you get competing interests. And when all these competing interests start fighting against one another, you end up in a stalemate. And it leads to a log jam with respect to the status quo.

LUCILA OHNO-MACHADO: Mm‑hmm. Ah, great. And essentially, the... intended use is an important part. Of, of all of this.

So, Mary Majumder asks: What are your views on solidarity ‑‑ suggesting "for all of us" ‑‑ in addition to or rather than altruism, which suggests "for others"?

EFFY VAYENA: I can say a few words; I'm sure Brad also has thoughts on this, but, um. One of the... policy themes and principles in our, and the work when we reviewed various frameworks, we saw that solidarity had gotten the least ‑‑ it was the one with, you know, almost no attention. So it's almost in the theoretical work that was done! Not recently. That was in actual ‑‑ not in scholarship. But in actual frameworks, solidarity wasn't featuring as something we need to pay attention to. And I think that's wrong? I think solidarity can, you know, bring out a lot. But! What was interesting is this development recently, the Lancet Commission that I mentioned, and the citation is in my slides, actually proposed a whole new concept: data solidarity. And it's, I think, what you're suggesting here, that if we think that solidarity ‑‑ we can support a solidarity for the data uses, that could get us out of some of those bottlenecks.

The only problem I have with solidarity ‑‑ not as a principle! At all. But rather, as how do we actually DO it? How do we DO solidarity? How do we... make people think that way, or act that way? And... what makes me a little, to question, not the value of it, but rather the enforcement and implementation, is when we look also at those people who sign consent forms and are asked whether they will contribute their data for research, most of them say yes! So, whether for altruism or solidarity seem to be okay with this. But then, somewhere, things get stuck! So, is that... would ‑‑ you know, making people feel more solidarity, or giving them ways to express their solidarity, get us out of it? I'm, I'm not, I'm not so sure. But as a value that we need to rely more and maybe figure out how to use it more effectively in our governance, I think that's certainly important.

BRADLEY MALIN: So, so... to pick up on this, there's a question about who are we asking solidarity in. And there's a difference between asking for solidarity in the individuals who are going to participate in a program, versus the organizations that believe that they have a responsibility to contribute information on their behalf. This is where things get a little tricky, in that I was ‑‑ I think I was alluding to this before, in that different scenarios constitute different control frameworks. So, for instance, in public health reporting, individuals can consent. This is just ‑‑ there's an expectation that that information is going to be collected and, and used. And in the event that you DO allow for consent, you potentially run into the problem of consent bias. In that you run into problems of trust, and that some subgroups may trust an expected user or potential user different than other groups, and then you end up with potential bias in the data.

And this is one of the reasons why this notion of like de‑identification or anonymization was brought to bear. (buzzing in background) Because it tried to remove in notion of consent bias. (perhaps a leafblower) So I just want to make sure that that gets laid on the table.

The second issue, though, is that, um... sometimes there is this notion of paternalism. In that, you have to make a, we have to figure out what's the right type of governance. Like, do we allow individuals to assert their control? Or do we say that there IS some higher level that gets to make decisions on when information gets to go out the door? I mean, this is like a more generalized view of what I was just talking about. But I think it's important to recognize that sometimes, you know, the goods of society and the goods of, the perspectives of society and the perspectives of individuals may be... in a little bit of conflict. And that's not necessarily a bad thing; it's just that there needs to be some type of shared decision‑making in these situations.

One other point I wanted to make ‑‑ I'm trying to look at all my notes while talking to you ‑‑ is that there's also this notion of individuals versus families! And since we are talking about, you know... an NHGRI‑sponsored environment, and we're talking about genomics, there's this question about who are even the individuals who should be at the table involved making decisions? So, for instance, if you have an individual that says they want to share their genomic record, we know that there's a potential for disclosing, or inferentially disclosing, information about family members! You know, whether it be an individual's parents, or their children, or their relatives. Their, you know, same‑point relatives like siblings. And so, you know, we don't go out and ask permission for everybody that you're related to, to determine if an individual can share information about themselves. And so there's, there's considered to be some, some limits to where, where this notion of, ah, an individual's autonomy kind of lives.

So, this is a really tough thing to wrestle with. Because, you know, at least in the U.S., we have like that really strong belief that individuals get to act on their own behalf. But in certain situations ‑‑ and I think we're particularly seeing this with Native American tribes, and with other subpopulations in the U.S. ‑‑ that there's an expectation of a different governance structure! That it's NOT the individual making decisions for themselves, but that there are group‑based decisions that need to be determined! And that can influence the governance structure, as well.

LUCILA OHNO-MACHADO: Thank you so much. And I just want to observe time and know that Mildred will do a closing. So perhaps we can take the last question? Which is from Ilaria Galasso: What are the specific concerns in terms of privacy and consent for socioeconomically disadvantaged individuals and communities?

BRADLEY MALIN: I feel like that's a question directed at you, Lucila. (chuckles)

LUCILA OHNO-MACHADO: Well, ah ‑‑ yeah, I have worked, ah, and continued to work with some... disadvantaged groups. And as was said, the governance is quite difference, and we've got to respect that. There is clear evidence from our work that, ah... some groups are less likely to consent for having their data in research. Which then creates this disparity. Downstream. So, it... it is a matter, I think, of being more transparent. And kind of transmitting, communicating, about the advantages of sharing. For society. And the risks, also being very clear about the risks, what they are and what we know about... events that, that happen because of breaches and so on. But. I think it's, again, I want to thank you both for providing a very balanced view about the pros and cons. And we could debate this for... quite a long time. You know. And also on the uses. Public health versus research of other kind. Versus individual care. Right? And individual benefits, in addition to individual risk.

So I apologize; we won't be able to go through all the questions. But you can see everyone appreciated the talks very much. And I'm sure over time we'll have... on the subsequent meeting, a chance to talk about those items.

So I wanted to get back to Mildred, 'cause I know she wants to close down the session. So, thank you both. Excellent talks and discussion. Not a debate; a discussion. Thank you.

MILDRED CHO: Thank you very much. I'm just gonna leave my camera off, 'cause my bandwidth, ah, limitations. But I do invite you to join us at, using the link in the chat, for our conversation that will be more informal, but we can continue to answer some of these questions afterwards. So please click on that to go to our conversations afterwards. And also, want to remind people ‑‑ because I get a lot of questions about this ‑‑ that this will be, this recording, and also all the wonderful resources in the chat, will be available on ELSIhub.org. So please go there to, to get all these articles and other references.

And please also join us next month in May for our next ELSI Friday Forum, which is called Addressing Algorithmic Harms: Practices and Provocations for Health AI. You'll receive a post‑event survey, which I'll encourage you to complete. So, to help us out to bring new topics and speakers to you at your suggestion. And I wish you a wonderful weekend. Please join us again for our, um, continuation of this conversation, using the link that's in the chat right now. Thank you so much.