Written Testimony of Atul Butte, MD, PhD
Priscilla Chan and Mark Zuckerberg Distinguished Professor, University of California, San Francisco, Director, Bakar Computational Health Sciences Institute and Chief Data Scientist, University of California Health

Committee on Energy and Commerce Subcommittee on Health

Hearing entitled "The Future of Biomedicine: Translating Biomedical Research into Personalized Health Care."

December 8, 2021

My name is Atul Butte. I am a physician scientist living in Northern California. My title at the University of California, San Francisco is that of Distinguished Professor and Director of the Bakar Computational Health Sciences Institute. I am also the Chief Data Scientist of the University of California Health System, across the 6 medical schools and health centers of the University of California. I am a member of the National Academy of Medicine. The views expressed here are my own, and do not represent the views of the University of California or any of the organizations discussed below.

I am here to provide testimony on the importance of data, data sharing, open access data, and clinical data, in the area of biomedical research, especially with respect to precision medicine.

As background, in 2013, I was proud to be recognized as a White House Champion of Change in Open Science for promoting science through the use of publicly available data. That year, White House Office of Science and Technology Policy Director John Holdren had directed Federal agencies with more than $100 million in R&D expenditures to develop plans to make the results of federally funded research freely available to the public. That directive led to a huge ongoing release of data generated by the government or funded by the government, to reach the public.

Indeed, our national data resources are growing, but we don't often think about this biomedical data infrastructure like we do think about other national resources, like our national labs, our national parks, our electrical grid, or our roads and bridges .

I will highlight a few examples here, all of which are housed at or through NIH.

Genbank is the repository containing all publicly available DNA sequences, for example those referenced in scientific publications or patents. Nearly 40 years old and now containing over 2 billion deposited genetic sequences, Genbank predates the invention of the World Wide Web, web browsers, or even CD-ROMs. But indeed, Genbank is still a relevant world-wide home for even SARS-CoV-2 sequences.

The Cancer Genome Atlas: one of the many disease-specific databases funded by the National Institutes of Health, specifically the National Cancer Institute. Launched in 2005 across many major institutions across America and the world, the program collected and studied over 20 thousand cancer samples across 33 types of cancer, and probed these tissues in many ways, studying their genes and proteins, and alsothe clinical aspects of the patients, in terms of the drugs successfully or unsuccessfully used, and even the imaging studies that captured the cancer, such as mammography or CT scans. Though the program has formally ended, the data is still there, publicly accessible, and can likely still fuel hundreds of research careers and companies.

PubChem: also housed at NIH, the world's largest collection of freely accessible chemical information, with 111 million compounds tested in over a million screening assays, resulting in a quarter billion bioactivities measured. It's probably a safe bet that many future drugs might just be sitting there as a column in this database, waiting to be discovered and developed.

PubMed Central: already 21 years old, with 7.5 million full text scientific articles readable by the public without charge. When it started, scientists like me would be asked to voluntarily submit our papers into this repository for access; now it is less of a choice and much more of a mandate. The implementation is not perfect and not all publications are accessible here in a timely manner, but at least members of the public can now access the published results of government-funded science for free.

All of Us research cohort: striving to recruit a million or more Americans, gathering their molecular and health data to enable researchers to discover and study predispositions to chronic diseases, how they are treated, and how they might be prevented.

ImmPort: working with the National Institute of Allergy and Infectious Diseases, my team in collaboration with Peraton works with researchers in allergy, immunology, inflammation, transplantation, and infectious diseases, including COVID-19, to help disseminate their data to the public.

These are just six out of many hundreds of NIH administered or funded data repositories, and out of literally tens of thousands of repositories run or funded by

others around the world. This emphasizes the volume and complexity of data needed to understand the human condition, and is still just the tip of the iceberg on all the data we will need to develop the next generation of cures and treatments.

And we are expecting much more data to come. Starting in January 2023, NIH will be executing on their new policy, requiring all NIH-supported research that generates scientific data to include a Data Management and Sharing Plan. It will be important to ensure these plans are good plans, evaluated closely, and enforced. If a research team gets funding because they are committing to disseminating their research data, then we should make sure that team ends up releasing their data.

It is easy to immediately think of open biomedical data as valueless, as we are used to seeing much content on the internet available at no charge. But instead, think of it this way: this openly accessible data is the raw research product of some of the best scientists in the world, funded by public research dollars. If you know why and how to access and use this data, and if it's submitted in a findable and usable way, then you essentially have these very best scientists working for you. I term this "retroactive crowd sourcing": these top scientists are out there on the internet helping you, and they don't even know they are helping you.

My own research lab has been using these large data assets to create novel diagnostics and therapeutics. A few have actually become startup companies. Three of my own examples:

In 2014, we analyzed publicly available molecular datasets around pregnancy complications, and developed a diagnostic for a specific complication called preeclampsia, which affects two hundred thousand women in the United States each year. We got the diagnostic to work in the lab, spun out a company called Carmenta, launched a prospective trial, and just as it was starting, the startup was acquired. From starting in the lab to acquisition of the spinout company took only about two years. The science continues in the acquiring company, which itself just had an IPO.

In 2011, we demonstrated how we could successfully use public molecular data to find new uses for drugs, called drug repositioning. With this experience, we launched a company called NuMedii, raised over $10 million, now pursuing drugs for fibrotic diseases with dozens of jobs created over these years.

And in 2008, we worked on the first patient to show up with a whole genome sequence, and built a huge data asset by manually reading and coding thousands of research publications in genetics, making a master list of the medical genome. With this work

and other similar databases, several of us launched a company called Personalis in 2011. The company raised over $70 million from private investors, and then had an initial purchase offering or IPO two years ago, now with around 200 employees and a market cap around a billion dollars. And continuing the data lifecycle, Personalis helps the Veterans Administration in the DNA sequencing used for its Million Veterans research Program.

In these three examples, creative use of data led to discoveries and inventions, which led to scientific publications, which leads to intellectual property, leading to the creation of companies, new jobs, and careers. I believe more biomedical researchers, especially academic faculty, should use data and computation to not just write more scientific publications, but to bring solutions to patients and doctors and families, and increasingly this means through entrepreneurship. And patients benefit, from earlier diagnoses to more available therapies, based on the increasingly specific and particular nature of their conditions.

A newer source of biomedical data is the data surrounding all the elements of clinical care. We in the United States have spent billions of dollars to acquire data on patients, through electronic health record systems. One of the most exciting roles I now have is Chief Data Scientist for the entire University of California Health system. Across our 6 medical schools, 12 hospitals, and hundreds of clinical centers, we've seen and treated over 7 million patients over the past 10 years. We've built a secure central data warehouse that we carefully use for operational improvement and promoting quality patient care, and, when deidentified, to enable the next generation of clinical research. Data from over 200 million encounters is saved, with over 560 million procedures, more than 760 million medication orders and prescriptions, and with over 2 billion vital signs measurements and test results. I call this data the most expensive in America. We pay doctors to type so much of this in. The narrative I want to make sure I leave you with is that given how much money we have collectively spent on getting this data, it will be a national tragedy if we don't use this data -- of course safely and responsibly – if we don't use this data to improve the practice of medicine.

Used respectfully and responsibly, with appropriate safeguards in place to protect the security and privacy of the data, this clinical data can inform patients as to the details of what's going on in their care, and what's next. My health system and many others now transmit data to patients through federal standards, but this is just the beginning, and a lot more effort is needed to expand these standards and their use. Data on the pricing of care and services can also help patients select the right level of care and an affordable price.

Clinical data can also help health systems and payers. We have already seen examples within the University of California where we can find and eliminate unnecessary use of specific medications. We can use these records to help patients switch therapies, from brand name to generic drugs. And we can compare the effectiveness of drugs, devices, and medical procedures, so that we can start to promote better, safer, and more cost effective therapies.

Collectively our nation's clinical data can document health equity, or inequity. However, clinical data is not tasked for this purpose right now. I am proud that Dr. Carrie Byington, our Executive Vice President for University of California Health, recently signed the Health Equity Pledge along with 40 other institutions to leverage our clinical data to document and address disparities.

Use of clinical data is just at the beginning, and it is helpful to see federal agencies seeing the value of this data, including the FDA which is now promoting programs in real-world evidence to speed the approval of certain drugs or indications. However, the science, engineering, and data infrastructure could still use a lot of investment. We ourselves see many dozens of new uses for clinical data, which we believe aregoing to lead to new science and entrepreneurship.

The data is there, to make sure the right drug, device, or procedure really should be used in the right way, to drive the safest, most effective, and highest value therapy for our patients, and all of this is just another way to just say "data-driven precision medicine."

But we must acknowledge there may be winners and losers in this new health data ecosystem. For example, data suggesting a particular drug should not be used does mean that the company making and marketing that drug would be adversely affected. Pharma companies, hospitals, and payers all compete in our health care system, and one of these having data on the products and services of another could lead to favorable and unfavorable shifts in utilization.

So we have to do better, by working with data in an open, fair, accountable, and governed way. And collecting more data isn't just the challenge. Making sure we collect data in a fair, responsible, and transparent way, and ensuring the data collected properly represents all our patients is of utmost importance. Imagine considering the purchase of a self-driving car that was only trained on roads in Mountain View, California. You would never accept such a car, that didn't know how to run in deep snow or blinding rain. So similarly, we should never utilize a self-driving medical algorithm trained only on a quarter of American patients. We should know and

document what's in the algorithms we're building, sharing, and buying, and ensure they are trained on data covering the diversity of Americans, and engineered by data scientists that cover the diversity of America.

Finally, it is essential to recognize that the security and privacy of this clinical data cannot be short-circuited or short-changed.

I will end with five specific recommendations.

First, we just do not have enough people trained to use our data. As more sophisticated datasets get generated, such as specialty clinical or molecular measurements, the need for data scientists with sophisticated training becomes more important. More funding should be made available for training at all stages, from college through post-graduate, teaching critical thinking, statistics, programming skills, database skills, design and visualization.

Second, federal funding for data repositories remains quite variable, and sometimes too arbitrary. A data repository that is used by thousands of researchers to drive enormous scientific and economic value might still have to justify its funding and existence every 5 years. And if defunded, too often that valuable data disappears.

Third, there are opportunities to open more Federal Government-related data to others. Imagine if the VA Million Veterans Program genetics data was available to more researchers. Imagine if the millions of chest x-ray images from federally-run hospitals and clinics was available to AI engineers, to build a novel tool to help read them, and a company around that. Of course, datasets like these can only be shared in carefully regulated ways. But are there ways to invest in technological solutions to make that access easier, to make better use of these national data resources?

Fourth, let's ensure the new 2023 NIH policies for data sharing do carry through, and that we create a culture that research data is disseminated with the public, whether molecular and cellular studies or clinical trials. More than ever, the public deserves to see more raw data behind our scientific findings, and allowing safe and responsible access to this raw data could help enhance trust between scientists and the public. This can help ensure our science is true and reproducible and robust, while preparing for a world of rogue analysts waiting to promote their own potentially incorrect studies and interpretation.

Fifth, we need to build on programs like the new NIH AIM-AHEAD (or at least ensure their funding continues), to not only make sure diversity is covered in biomedical data sets, but diversity is promoted and enhanced among the data scientists themselves.