COLUMBIA PRECISION MEDICINE INITIATIVE

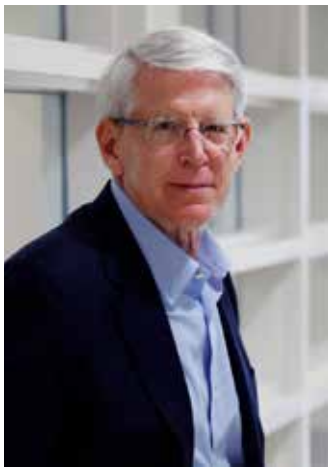
ADVANCES IN PRECISION MEDICINE

HARMONIZING CLINICAL AND GENOMIC DATA

FOURTH ANNUAL CONFERENCE APRIL 24, 2020

COLUMBIA | PRECISION MEDICINE

Welcome Letter



I am delighted to welcome you to the Fourth Annual Columbia Precision Medicine Initiative (CPMI) conference, Advances in Precision Medicine: Harmonizing Clinical and Genomic Data.

Due to the COVID-19 pandemic, this year's conference is taking place online.

In our previous annual CPMI conferences, we have considered the impact of genetics on clinical care, cancer genomics, and big data on precision medicine. The most significant challenge facing the effective practice of precision medicine is the problem of managing and understanding massive amounts of clinical and genomic data generated daily through regular patient care. How can multiple types of data be integrated with electronic health records and analyzed to effectively inform clinical care? We have been fortunate to assemble an extraordinary group of speakers to address this question.

I sincerely hope that you enjoy, and are informed by, the conference.

Tom Maniatis, PhD
Director, Columbia University Precision Medicine Initiative
Isidore S. Edelman Professor of Biochemistry and Molecular Biophysics

Conference Schedule

9:50 a.m.	**Tom Maniatis, PhD:** Welcome
9:55 a.m.	**David Goldstein, PhD:** Introduction
10:00 a.m.	**David Ledbetter, PhD, FACMG,** Geisinger
10:35 a.m.	**Dan Roden, MD,** Vanderbilt University Medical Center
11:10 a.m.	**Marylyn D. Ritchie, PhD,** University of Pennsylvania Perelman School of Medicine
11.45 a.m.	**Rex L. Chisholm, PhD,** Northwestern University Feinberg School of Medicine
12:20 p.m.	Lunch break
12:55 p.m.	**Chunhua Weng, PhD, FACMI,** Columbia University
1:30 p.m.	**Nancy J. Cox, PhD,** Vanderbilt University Medical Center
2:05 p.m.	**David Madigan, PhD,** Columbia University
2:40 p.m.	**Nicholas P. Tatonetti, PhD,** Columbia University

David Goldstein, PhD

John E. Borne Professor of Genetics and Development
Director, Institute for Genomic Medicine at Columbia University Irving Medical Center

David Goldstein, PhD, is the John E. Borne Professor of Genetics and Development and director of the Institute for Genomic Medicine at Columbia University Irving Medical Center. Previously, he was a professor of genetics at Duke University and the director for the Center for Human Genome Variation. He trained in theoretical population genetics at Stanford University. Dr. Goldstein's primary research interests include human genetic diversity, the genetics of disease, and pharmacogenetics. The Goldstein group and collaborators have discovered a number of disease-causing genes and syndromes, in particular neurological and infectious diseases. In addition, Dr. Goldstein and colleagues are interested in identifying molecular (genetic) pathways that drive the appearance and progression of distinct chronic kidney disease (CKD) subtypes and genetic drivers of CKD, with a focus on IgAN and FSGS. Dr. Goldstein was elected a fellow of the American Association for the Advancement of Science (AAAS) in 2013, received the University of North Carolina at Chapel Hill Institute for Pharmacogenomics and Individualized Therapy (IPIT) award for clinical services in 2012, and was a recipient of one of the first seven nationally awarded Royal Society/Wolfson research merit awards in the United Kingdom for his work in human population genetics. In 2013, Dr. Goldstein chaired the Gordon Research Conference in Human Genetics, and he is currently serving on the Advisory Council at the National Institute of Neurological Disorders and Stroke at the National Institutes of Health (NIH) and as chief genomics adviser to AstraZeneca.

David Ledbetter, PhD, FACMG

Geisinger

David Ledbetter is executive vice president and chief scientific officer at Geisinger. Previously, he held academic and leadership positions at Emory University, the University of Chicago, and the National Center for Human Genome Research at NIH. He is a graduate of Tulane University and earned his PhD at the University of Texas at Austin. He is an internationally recognized expert in genomics and precision medicine, having focused his early research efforts on discovering the genetic causes of childhood neurodevelopmental disorders such as autism, and the translation of new genomics technologies into clinically useful genetic tests for early diagnosis and intervention. His current research interests include leveraging longitudinal electronic health information with large-scale DNA sequencing to determine the clinical utility and cost-effectiveness of precision medicine approaches in a real-world health system setting.

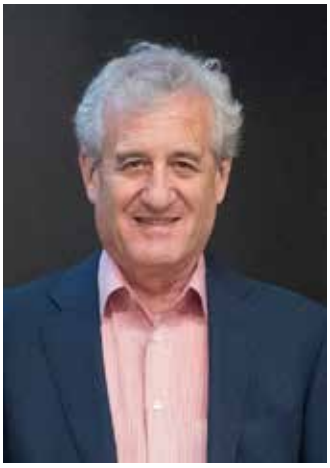
Geisinger's Genomics and Precision Health Initiative

ABSTRACT:

Geisinger initiated MyCode® in 2007 to link longitudinal electronic health records (EHR) and other clinical data to DNA sequence and genotype data. We have consented >260,000 patient-participants and, in partnership with the Regeneron Genetics Center, have exome sequence and genome-wide SNP data for 145,000 individuals. To date, analysis of exome data has been completed in 65,012 eligible MyCode participants with 1,497 (2.3%) receiving clinically confirmed results in the medically actionable genes from the American College of Medical Genetics and Genomics (ACMG) secondary findings list plus the HFE and CFTR genes. These results have generated high-level engagement from our clinicians and patient-participants, which has fueled interest in implementing DNA screening as part of routine clinical care.

We have launched clinical exome sequencing over the last year in two large primary care clinics. This screening test is offered to adult patients by their primary care clinician with verbal consent, irrespective of personal or family history of disease. All patients who undergo the clinical DNA screen receive a laboratory report within a timely manner, regardless of whether the screen is positive or negative. The ordering provider discloses negative results, while positive results are disclosed by a genetic counselor once the ordering provider is notified.

Overall, our results show that clinic-based DNA screening has a screen positive rate consistent with previous research studies. In addition, although many of our screen positive patients had a personal or family history consistent with the genomic finding, most did not have previous clinical genetic testing, demonstrating the potential for DNA screening to detect patients who would otherwise not be offered diagnostic genetic testing.



Dan Roden, MD

Vanderbilt University Medical Center

Dan Roden is professor of medicine, pharmacology, and biomedical informatics and senior vice president for personalized medicine at Vanderbilt University Medical Center. He grew up in Montreal and received his medical degree and trained in internal medicine at McGill University. He then went to Vanderbilt, where, after fellowships in clinical pharmacology and cardiology, he joined the faculty and has remained there ever since. His research program studies how genetic variation affects human disease susceptibility, with a focus on pharmacogenetics, and on genetic determinants of abnormal heart rhythms, especially those induced by drugs. After serving as chief of the Division of Clinical Pharmacology for 12 years, Dr. Roden was tasked in 2006 with leading Vanderbilt's efforts in personalized medicine. Under his leadership, Vanderbilt has become internationally recognized for cutting-edge programs in this area, including the large (250,000 sample) biobank BioVU and the EHR-based preemptive pharmacogenetic program PREDICT. He has been principal investigator for the Vanderbilt sites of the Pharmacogenomics Research Network since 2001 and of the Electronic Medical Records and Genomics (eMERGE) Network since 2007.

Biobanks, Electronic Health Records, and Precision Medicine

ABSTRACT:
This talk will describe how the development of large sets of electronic health records (EHRs) linked to genomic information has been an engine for discovery and methods development, and how EHRs will be used to deliver a vision of genomically informed medical care.

Early studies, at our center and across the eMERGE Network, used genome-wide association studies (GWAS) to replicate known relationships between common diseases or drug responses identified in EHRs and genomic variants, and to discover new relationships. Initial lessons have included development of robust methods for electronic phenotyping and an increasing appreciation of the need for very large data sets in order to be able to study subsets with complex phenotypes such as drug response or longitudinal disease evolution.

The growth of EHRs linked to DNA data has also propelled the development of new analytic approaches to understand genotype-phenotype relations. For example, associations between specific genetic variants, or collections of genetic variants, and the broad range of disease-associated human phenotype(s) in EHRs can be interrogated using the "phenome-wide association study" (PheWAS). PheWAS has been used to validate GWAS signals and to discover pleiotropic gene effects, and it holds promise in disease subsetting, drug development and repurposing, and biomarker identification and validation. Analysis of EHR-derived phenotypes using "phenotype risk scores" shows promise in identifying patients with unrecognized genetic diseases.

It is a common vision that genomic information will be routinely used in the care of patients. It is almost inconceivable that large-scale implementation of this vision could be executed in the absence of EHRs that provide advanced clinical decision support to deliver real-time advice to practitioners based on genetic or phenotypic information in an individual EHR.

Marylyn D. Ritchie, PhD

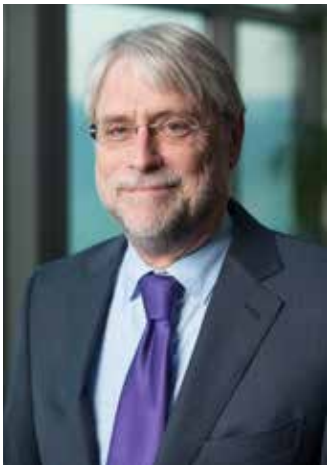
University of Pennsylvania Perelman School of Medicine

Marylyn D. Ritchie, PhD, is a professor in the Department of Genetics, director of the Center for Translational Bioinformatics, and associate director for bioinformatics in the Institute for Biomedical Informatics at the University of Pennsylvania Perelman School of Medicine. She is also associate director for the Penn Center for Precision Medicine. Dr. Ritchie is a translational bioinformatics scientist, biomedical informatician, and computational human geneticist with a focus on developing novel approaches for understanding the relationship between our genome and human phenotypes. She has expertise in developing novel bioinformatics tools for complex analysis of big data in genetics, genomics, and clinical databases, in particular in the area of pharmacogenomics. Dr. Ritchie has over 15 years of experience in the analysis of complex data and has authored over 300 publications. She has received several awards and honors including selection as a Genome Technology Rising Young Investigator in 2006, an Alfred P. Sloan Research Fellow in 2010, and a KAVLI Frontiers of Science Fellow by the National Academy of Science from 2011 to 2014, and she was named one of the most highly cited researchers in her field by Thomas Reuters in 2014. Dr. Ritchie also won First Place in the AMIA "Why Informatics" Video Contest in 2017. Dr. Ritchie has extensive experience in all aspects of genetic epidemiology and translational bioinformatics as it relates to human genomics. She also has extensive expertise in dealing with big data and complex analysis including GWAS, next-generation sequencing, data integration of meta-dimensional omics data, Phenome-wide Association Studies (PheWAS), and development of data visualization approaches.

The Power of Medical Biobanks for Precision Medicine

ABSTRACT:

Biomedical data science has experienced an explosion of new data over the past decade. Abundant genetic and genomic data are increasingly available in large, diverse data sets due to the maturation of modern molecular technologies. Along with these molecular data, dense, rich phenotypic data are also available on comprehensive clinical data sets from health care provider organizations, clinical trials, population health registries, and epidemiologic studies. The methods and approaches for interrogating these large genetic/genomic and clinical data sets continue to evolve rapidly, as our understanding of the questions and challenges continues to develop. Through applying bioinformatics, statistics, and machine learning approaches to the rich phenotypic data of the EHR, these data can be mined to identify new and interesting patterns of disease expression and relationships. We have been exploring various translational bioinformatics technologies for evaluating the phenomic landscape to improve our understanding of complex traits. These techniques show great promise for the future of precision medicine.



Rex L. Chisholm, PhD

Northwestern University Feinberg School of Medicine

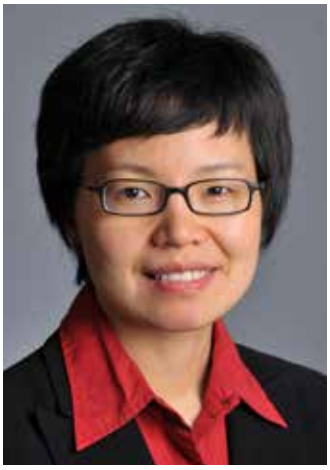
Rex Chisholm is the Adam and Richard T. Lind Professor of Medical Genetics in the Feinberg School of Medicine and professor of cell and developmental biology and surgery. He was the founding director of the Center for Genetic Medicine. Since 2007, he has served as vice dean for scientific affairs in the Feinberg School. In October 2012, he was also appointed associate vice president for research at Northwestern University.

As vice dean for scientific affairs, Dr. Chisholm is responsible for research space in Feinberg, research core facilities and the broader research environment. He also oversees the PhD and MS training programs in Feinberg. As associate VP for research, he has oversight responsibilities for the Office for Sponsored Research and the Center for Comparative Medicine, as well as several University Research Centers.

Precision Medicine at Northwestern: Discovery and Implementation in the eMERGE Network

ABSTRACT

Northwestern Medicine established the NUgene Biobank in 2002. This fully-consented biobank has served as the impetus to integrate phenotypes derived from Electronic Health Records with genomic information for the purpose of making new disease gene associations, using genomic information to drive therapeutic selection and implementation of precision medicine. Initially we demonstrated that in the case of type 2 diabetes (T2D), a biobank of "random" participants could provide T2D cases and controls that replicated gene associations made in purpose-built cohorts. Since then, in partnership with the Electronic Medical Records and Genomics consortium (eMERGE), we have implemented more than forty phenotypes. Northwestern has also been a leader in developing strategies for associating genomic information with the EHR through the use of an ancillary genomics system. This allows the health system to store only clinically relevant genetic variation while providing an associated means of storing and reassessing clinical relevance of genomic variation as evidence matures. Finally, we will describe our recent efforts at measuring genetic variation in the list of 59 genes the American College of Medical Genetics and Genomics recommends, returning pathogenic variation even in the absence of clinical indication. Understanding the value of this returned information is an important prerequisite for larger-scale implementation of precision medicine.



Chunhua Weng, PhD, FACMI

Columbia University

Chunhua Weng is a tenured professor of biomedical informatics at Columbia University and an elected fellow of the American College of Medical Informatics. She also co-leads the Biomedical Informatics Resource for Columbia's Irving Institute for Clinical and Translational Science. Dr. Weng holds a PhD in biomedical and health informatics from University of Washington at Seattle. She has been an active researcher in the field of clinical research informatics since 2000 and has published extensively on text knowledge engineering for clinical research eligibility criteria, EHR data quality assessment and data analytics, and high-throughput EHR phenotyping.

Deep Phenotyping on EHR Narratives Facilitates Genetic Diagnosis by Clinical Exomes

ABSTRACT

Integration of detailed phenotypic information with genetic data is well established to facilitate accurate diagnosis of hereditary disorders. As a rich source of phenotypic information, electronic health records (EHRs) have the potential to empower diagnostic variant interpretation. However, how to accurately and efficiently extract phenotypes from heterogeneous EHR narratives remains a challenge. This talk describes a high-throughput EHR phenotype extraction and analysis framework that performs human phenotype ontology (HPO) concept extraction and normalization from EHR narratives and prioritizes disease genes based on the HPO-coded phenotypic manifestations. Our results on four retrospective cohorts show the promise of leveraging EHR data to automate phenotype-driven analysis of clinical exomes or genomes, facilitating the implementation of genomic medicine on scale.

Nancy J. Cox, PhD

Vanderbilt University Medical Center

Nancy J. Cox is the Mary Phillips Edmonds Gray Professor of Genetics and director of the Vanderbilt Genetics Institute and Division of Genetic Medicine at Vanderbilt University Medical Center. She is a quantitative geneticist with a long-standing research program in identifying the genetic component to common diseases and related traits, including pharmacogenomics. Current research is focused on developing approaches for integrating genome variation with genome function and the medical phenome using BioVU, Vanderbilt's biobank, with DNA samples on >285,000 subjects linked to high-quality electronic health records going back 30 years, augmented by phenome data in another 2.8 million subjects. Dr. Cox earned a BS in biology from the University of Notre Dame in 1978, a PhD in human genetics from Yale University in 1982, and did postdoctoral research at Washington University and the University of Pennsylvania, before spending 28 years on the faculty at the University of Chicago, and then joined Vanderbilt in 2015.

Exploring and Exploiting Pleiotropy in EHR-Linked Biobanks

ABSTRACT

Among the unique advantages of biobanks linked to electronic health records (EHR) is the ability to study genetic pleiotropy at scale. I describe here recent advances in prediction modeling for human gene transcript levels that incorporate information on tissue correlations in gene expression and DHS to improve prediction substantially. Application of such models to BioVU, Vanderbilt's biobank, with DNA samples on >285,000 subjects linked to high-quality EHR going back an average of 10–15 years and more than 30 years in some subjects, allows us to assess the association of the medical phenome to the natural variation in human gene expression. We describe here how these data can be used to study genetic pleiotropy, including underlying mechanisms, as well as how pleiotropy might be exploited to develop larger targets for discovery and richer avenues for drug development and prevention of later-onset diseases.

David Madigan, PhD

Columbia University

David Madigan is professor of statistics at Columbia University, where he recently concluded a term as dean of Arts and Sciences. He received a bachelor's degree in mathematical sciences and a PhD in statistics, both from Trinity College Dublin. He has previously worked for AT&T Inc., Soliloquy Inc., the University of Washington, Rutgers University, and SkillSoft, Inc. He has more than 200 publications in such areas as Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance, and probabilistic graphical models. He is an elected fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. He has served terms as editor-in-chief of *Statistical Science* and of *Statistical Analysis and Data Mining: The ASA Data Science Journal*.

Towards Honest Inference from Real-World Health Care Data

ABSTRACT

In practice, our learning health care system relies primarily on observational studies generating one effect estimate at a time using customized study designs with unknown operating characteristics and publishing—or not—one estimate at a time. When we investigate the distribution of estimates that this process has produced, we see clear evidence of its shortcomings, including an apparent overabundance of statistically significant effects. We propose a standardized process for performing observational research that can be evaluated, calibrated, and applied at scale to generate a more reliable and complete evidence base than previously possible. We demonstrate this new paradigm by generating evidence about all pairwise comparisons of 39 treatments for hypertension for a relevant set of 58 health outcomes using nine large-scale health record databases from four countries. In total, we estimate 1.3M hazard ratios, each using a comparative effectiveness study design and propensity score stratification on par with current one-off observational studies in the literature. Moreover, the process enables us to employ negative and positive controls to evaluate and calibrate estimates ensuring, for example, that the 95% confidence interval includes the true effect size 95% of the time. The result set consistently reflects current established knowledge where known, and its distribution shows no evidence of the faults of the current process.

This joint work is done with George Hripcsak, Patrick Ryan, Martijn Schuemie, and Marc Suchard.

Nicholas P. Tatonetti, PhD

Columbia University

Nicholas P. Tatonetti is associate professor of biomedical informatics at Columbia University, with interdisciplinary appointments in the Department of Systems Biology and the Department of Medicine. He received his PhD in biomedical informatics from Stanford University in 2012, and dual BS degrees from Arizona State University in computational mathematics and molecular biosciences/biotechnology in 2008. Dr. Tatonetti's research is focused on advancing our understanding of drug effects and drug combinations through the integration of observational clinical data and high-throughput molecular data. A recognized expert in adverse drug effects, Dr. Tatonetti is responsible for discovering previously unexpected drug interactions causing heart arrhythmias and glucose dysregulation.

Dr. Tatonetti is an Irving Scholar, a Kavli Fellow, the recipient of New Investigator Awards from the American Medical Informatics Association (2016) and the PhRMA Foundation (2014), and has been awarded more than $4 million in research funding. His work has received multiple awards in informatics and data science (2010, 2011, 2012, 2015, and 2016). Dr. Tatonetti's career has been profiled by *Science Magazine* (2011), Genome Web (2012), and AMIA (2016). He is author of more than 90 peer-reviewed scientific publications, inventor of two patents, and his work as been covered by the popular and scientific press, including *The New York Times, The Chicago Tribune,* and *The Boston Globe*.

Large-Scale Application of Machine Learning of Observational Data Reveals Disease Mechanisms and Genetic Associations

ABSTRACT

Data is transforming the scientific method across many domains. In drug safety, data from electronic health records and search logs are now being collected alongside traditional sources like case studies, spontaneous reporting systems, and model systems. These data sources present new opportunities for studying the phenomenological and molecular effects of active small molecules. However, they also present new challenges in data integration, statistical analysis, and the nature of hypothesis generation. I will discuss these opportunities and challenges and their role in the future of drug safety science.