June 6, 2022 | 8:30 a.m. – 5:00 p.m.

Symposium on Risks & Opportunities of AI in Clinical Drug Development

Pfizer | Northeastern University | ASA American Statistical Association Promoting the Practice and Profession of Statistics | Columbia University | Northeastern University Observational Health Data Sciences and Informatics Center

The Pfizer/ Northeastern/ASA/Columbia Symposium on Risks and Opportunities of AI in Clinical Drug Development is an event jointly sponsored by Pfizer Inc., The Roux Institute at Northeastern University, the American Statistical Association (ASA), and Columbia University.

Our world increasingly relies on data and computing to create knowledge, to make critical decisions, and to better predict the future. Data science has emerged to support these data-driven activities by integrating and developing ideas, concepts, and tools from computer science, engineering, information science, statistics, and domain fields. Data science now drives fields as diverse as biology, astronomy, material science, political science, and medicine–not to mention vast tracts of the global economy, key government activities, and quotidian social and societal functions.

The pharmaceutical enterprise has been slower to respond, especially to the rapid developments in AI, but tectonic shifts are underway in approaches to the discovery, development, evaluation, registration, monitoring, and marketing of medicines for the benefit of patients and the health of the community.

While there is much discussion about the potential of AI and modern machine learning tools to transform the drug development paradigm, there is a growing recognition of the paucity of research about the inevitable pitfalls and unintended consequences of the digital revolution in this important area of application. As we move toward personalized and truly evidence-based medicine, the use of AI and machine learning to optimize drug deployment raises a whole different set of challenges.

This forum is, therefore, expected to serve as a platform for distinguished statisticians, data scientists, regulators, and other professionals to address the challenges and opportunities of AI in pharmaceutical medicine; to foster collaboration among industry, academia, regulatory agencies, and professional associations; and to propose recommendations with policy implications for proper implementation of AI in promoting public health.

Keynote Speaker

M. Khair ElZarrad, PhD, MPH, Director, Office of Medical Policy, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Speakers and Panelists

- Javier Cabrera, Rutgers University
- Usama Fayyad, Northeastern University
- Asieh Golozar, Odysseus Data Services
- Shameer Khader AstraZeneca
- Kannan Natarajan, Pfizer Inc.
- Adler Perotte, Columbia University
- Ravi Parikh, University of Pennsylvania
- Anthony Philippakis, Broad Institute
- Prasanna Rao, Pfizer Inc.
- David Sontag, MIT
- Robert D. Truog, Harvard Medical School
- Mark J. van der Laan, University of California, Berkeley
- Robert Vandersluis, GSK
- Li Wang, AbbVie
- Marinka Zitnik, Harvard University

Registration Fees

Regular In-Person	$125
Regular Virtual	$75
Student	$0

In-person registration is all-inclusive of refreshments; breakfast and lunch will be served.

Program Co-Chairs: Demissie Alemayehu, Pfizer Inc. and David Madigan, Northeastern University

7:30 AM – 8:30 AM	Registration and Breakfast
8:30 AM – 8:35 AM	**Welcome – David Madigan*, Northeastern University**
8:35 AM – 9:25 AM	**Keynote Address – Chair, Kannan Natarajan*, Pfizer Inc.** *Responsive Regulations in the Age of Exponential Technologies.* M. Khair ElZarrad*, CDER, U.S. FDA
9:25 AM – 11:05 AM	**Plenary Session I – Chair, Demissie Alemayehu*, Pfizer Inc.** • *Data Science and Common Diseases.* Anthony Philippakis*, Broad Institute (25 min) • *ML and AI in Clinical Development and Statistical Innovation.* Li Wang*, AbbVie (25 min) • *Practical Approaches for Preprocessing Complex and High Dimensional Healthcare Data in Predictive Modeling Using Machine Learning Techniques, With Applications to Covid 19 Mortality and Wearable Device Data.* Javier Cabrera*, Rutgers University (25 min) • *AI in Health: From Digital Health to Transforming Life Sciences – Data and Method Considerations.* Usama Fayyad*, Northeastern University (25 min)
11:05 AM – 11:20 AM	Break
11:20 AM – 12:35 PM	**Plenary Session II – Chair, David Madigan*, Northeastern University** • *Finding Natural Experiments in Observational Data.* Adler Perotte, Columbia University (25 min) • *Framework for Assessing the Reproducibility of Observational Comparative Effectiveness Research.* Asieh Golozar*, Global Head of Data Science, Odysseus Data Services, Inc. (25 min) • *Towards Integration of Targeted Learning and Causal Inference in Drug Approval Process and Safety Analysis.* Mark van der Laan, University of California, Berkeley (25 min)
12:35 PM –1:30 PM	Lunch
1:30 PM – 2:45 PM	**Plenary Session III – Chair, Javier Cabrera*, Rutgers University** • *Augmenting Drug Development Using Clinical Trial Data Mining and Machine Intelligence.* Shameer Khader*, AstraZeneca (25 min) • *AI-Driven Clinical Documentation with MedKnowts for Cheap and Scalable Comparative Effectiveness and Safety Monitoring Studies.* David Sontag*, MIT (25 min) • *The Efficacy Arm In Silico: Opportunities and Challenges with Using Real-World Data in Drug Development and Comparative Effectiveness.* Ravi B. Parikh, U Pennsylvania (25 min)
2:45 PM –3:00 PM	Break
3:00 PM – 4:30 PM	**Panel Discussion – Ethical Issues with AI in Clinical Drug Development** **Moderator, Marinka Zitnik*, Harvard University** • Usama Fayyad*, Northeastern University • Prasanna Rao*, Pfizer Inc. • Robert D. Truog*, Harvard Medical School • Robert Vandersluis*, GlaxoSmithKline (GSK)
4:30 – 4:35 PM	**Closing Remarks – Demissie Alemayehu*, Pfizer Inc.**
5:00 PM	**Networking Reception**

**Designates an in-person speaker*

SPEAKER BIOS

Javier Cabrera, PhD

Javier Cabrera is a Professor in the Department of Statistics and the Department of Medicine, Rutgers University, and a member of the Cardiovascular Institute of New Jersey and the Institute of Quantitative Biomedicine. He is a winner of the 2010 SPAIG award of the American Statistical Association, a Fulbright fellow, and a Henry Rutgers fellow. He was Director of the Institute of Biostatistics at Rutgers University and the chief co-editor of the journal, Computational Statistics and Data Analysis. Professor Cabrera has numerous publications and books in Statistics and Biostatistics on diverse topics, including, Big Data for medical research, functional genomics, analysis of genomic data, statistical computing, graphics, and computer vision. He received his PhD from Princeton University.

M. Khair ElZarrad, PhD, MPH

M. Khair ElZarrad is the Director of the Office of Medical Policy (OMP) at FDA's Center for Drug Evaluation and Research (CDER), where he leads the development, coordination, and implementation of medical policy programs and strategic initiatives. Dr. ElZarrad currently leads multiple projects focused on exploring the potential utility of real-world evidence, innovative clinical trial designs, and the integration of technological advances in pharmaceutical development. Dr. ElZarrad is the rapporteur for the International Council for Harmonisation's ongoing work to revise the international Good Clinical Practice Guideline (ICH-E6(R2)). Prior to joining the FDA, he served as Acting Director of the Clinical and Healthcare Research Policy Division with the Office of Science Policy at the National Institutes of Health (NIH). At NIH he worked on policies related to human subject protections; the design, conduct, and oversight of clinical research; and enhancing quality assurance programs at pharmaceutical development and production facilities. He earned a doctoral degree in medical sciences with a focus on cancer metastases from the University of South Alabama, as well as a master's degree in public health from the Johns Hopkins Bloomberg School of Public Health.

Usama Fayyad, PhD

Usama Fayyad is the Inaugural Executive Director for the Institute of Experiential AI, Northeastern University and a Professor of the Practice in Northeastern University's Khoury College of Computer Science. Dr. Fayyad also serves as Chairman of Open Insights, a technology and consulting firm he founded in 2008 after leaving Yahoo. Leveraging open-source and Big Data technology with strategic consulting, Open Insights deploys data-driven solutions to grow revenue from data assets through Big Data strategy, new business models, data science, and AI/ML solutions.

Dr. Fayyad's storied career includes prior executive roles at OODA Health, Inc, Barclays Bank in London, Oasis500 (an appointment by King Abdullah II of Jordan) and several startups, including Blue Kangaroo Corp, DMX Group, and Digimine Inc. Most notably, he was the first person to hold the Chief Data Officer title when Yahoo acquired his second startup in 2004. At Yahoo, he built the Strategic Data Solutions group and founded Yahoo Research Labs. Dr. Fayyad also held leadership roles at Microsoft (1996-2000) and founded the Machine Learning Systems group at NASA's Jet Propulsion Laboratory (1989-2005). He has published over 100 technical articles on data mining, data science, AI/ML, and databases. Dr. Fayyad holds over 20 patents and is a Fellow of both the Association for the Advancement of Artificial Intelligence (AAAI) and the Association for Computing Machinery (ACM). He earned his doctorate in Engineering in Artificial Intelligence/Machine Learning from the University of Michigan, Ann Arbor. He has edited two influential books on data mining/data science and served as Founding Editor-in-Chief on two key journals. To learn more about Dr. Fayyad's extensive background in AI and Machine Learning, visit https://roux.northeastern.edu/people/usama-fayyad/.

Asieh Golozar, PhD

Asieh Golozar is the VP, Global Head of Data Science at Odysseus Data Services, Inc, leading a team of data scientists, epidemiologists and bioinformaticians focusing on epidemiological research and advanced and innovative analytics across a large global network of observational data. She also holds an adjunct position at the Johns Hopkins University (JHU) Bloomberg School of Public Health.

Dr. Golozar has more than 15 years of experience in life science research and medicine in industry, academia and government settings. She received her Doctoral degree in Epidemiology and a Master of Health Sciences in Biostatistics from JHU Bloomberg School of Public Health which was supported by a postdoctoral research fellowship award with the National Cancer Institute (NCI), Division of Cancer Epidemiology and Genetics. Prior to that, Dr. Golozar trained as a medical doctor at Tehran University of Medical Sciences. Upon receiving her PhD, joined the faculty at the Department of Epidemiology, JHU School of Public Health focusing on cancer and diabetes epidemiology and applying evidence-based findings to strengthen public heath infrastructure and policies. Dr. Golozar then joined the pharmaceutical industry where she worked as pharmacoepidemiology therapeutic area lead and expert at Regeneron Pharmaceuticals, AstraZeneca and Bayer and led and contributed to the integration of effective and efficient observational research strategy into the research and development, clinical development and life cycle management in different therapeutic areas specifically oncology.

Shameer Khader, PhD

Shameer Khader is a Senior Director of Data Science and Artificial Intelligence at AstraZeneca, USA. He leads a global team that leverages trans-disciplinary (biomedical, healthcare, and clinical) big data and machine intelligence to accelerate drug discovery and development. He has more than a decade of experience building and leading bioinformatics and data science in both academia and industry. He obtained his PhD in Computational Biology from the National Center for Biological Sciences, Bangalore, India. He completed his post-doctoral training in computation genomics and precision medicine from Mayo Clinic, Rochester, MN. He has published more than 140 peer-reviewed research publications, conference papers and patents in healthcare data science, bioinformatics, drug discovery, and precision medicine. His work was featured in media outlets including Forbes, Fast Company, Bloomberg News, and Times of India. He received multiple awards for his research contributions; His work on developing an open catalogue of drug repositioning has won the Swiss Institute of Bioinformatics' Bioinformatics Resource Innovation Award in 2017. Recently, he was recognized as one of the 100 Artificial Intelligence Leaders in Drug Discovery & Healthcare (DKI Global and Forbes). His TEDx Talk on Saving Lives Using Biomedical Data Science is available from https://www.youtube.com/watch?v=ZM4u4XIhm08

Kannan Natarajan, PhD

Kannan Natarajan is the Head of Global Biometrics and Data Management and is part of the Global Product Development Leadership Team at Pfizer Inc. The GBDM organization supports the global clinical development strategy and data sciences across all of Pfizer product portfolio. He is also the Chief Statistical Officer of Pfizer, managing statistical functional excellence across all Pfizer business units. Prior to joining Pfizer, Dr. Natarajan was Senior Vice President and Global Head of Oncology Biometrics and Data Management at Novartis Pharmaceuticals. He has been in the pharmaceutical industry for over 20 years working across various therapeutic areas. Dr. Natarajan holds a PhD in Statistics from the University of Florida.

Ravi B. Parikh, MD, MPP, FACP

Ravi B. Parikh is an Assistant Professor in the Department of Medical Ethics and Health Policy and Medicine at the University of Pennsylvania, Staff Physician at the Corporal Michael J. Crescenz VA Medical Center, and Senior Fellow at the Leonard Davis Institute for Health Economics. Dr. Parikh is a practicing oncologist who uses machine learning, quasi-experimental, and pragmatic clinical trial methods to study broad questions in cancer and advanced illness care. Dr. Parikh's work on medical technology and advanced illness has been published in Science, The New England Journal of Medicine, JAMA, and other high-impact publications. He is Senior Clinical Advisor at the Coalition to Transform Advanced Care (C-TAC) and sits on the Leadership Consortium of

the National Quality Forum. Dr. Parikh is a graduate of Harvard Medical School, Harvard College, and the John F. Kennedy School of Government. He completed a residency in internal medicine at Brigham and Women's Hospital and a fellowship in Hematology and Oncology at the University of Pennsylvania.

Adler Perotte, PhD

Adler Perotte is an Assistant Professor in the Department of Biomedical Informatics at Columbia University as well as the Chief Science and Innovation Officer at Spiden AG. Dr. Perotte's primary research areas are Bayesian inference and prediction based on electronic health record data and metabolomics. Prior to joining the department in this capacity, Dr. Perotte was an NLM Postdoctoral Fellow in Biomedical Informatics at Columbia University. Before that, Dr. Perotte worked as a Research Specialist between Princeton's Computational Memory and Artificial Intelligence laboratories studying probabilistic models of memory. While earning his MD from Columbia University, Dr. Perotte interned with Columbia's technology transfer office to evaluate promising university technology for commercial potential.

Anthony Philippakis, MD, PhD

Anthony Philippakis is the Chief Data Officer of the Broad Institute of MIT and Harvard, and the co-director of the Eric and Wendy Schmidt Center.

He trained as a cardiologist at Brigham and Women's Hospital, with a focus on rare genetic cardiovascular diseases. At the Broad Institute he is the founding director of the Data Sciences Platform, an organization of over 200 software engineers and computational biologists that develops software for analyzing genomic and clinical data. In addition to his roles at the Broad Institute and Brigham and Women's Hospital, Dr. Philippakis is a venture partner at GV, focusing on machine learning, distributed computing, and genomics.

Dr. Philippakis received his MD from Harvard Medical School and completed a PhD in biophysics at Harvard. As an undergraduate, he studied mathematics at Yale University, and later completed the Part III (equivalent to MPhil) in mathematics at Cambridge University.

Prasanna Rao

Prasanna Rao is an AI practitioner and industry thought leader whose current role is Senior Director, Global Head of AI/ML for Data Management and Monitoring at Pfizer Inc. He has over 30 years of experience in Information Technology and Analytics, with over ten years in Healthcare and Life Sciences. In his previous role as a Watson Solution Architect at IBM, he was instrumental in deploying many different AI systems, from idea to implementation, with various clients. In his current role, he works with diverse stakeholders, including machine learning developers and data scientists, to deliver innovation and drive adoption of AI.

David Sontag, PhD

David Sontag is an Associate Professor in the Department of Electrical Engineering and Computer Science (EECS) at MIT, and member of the Institute for Medical Engineering and Science (IMES) and the Computer Science and Artificial Intelligence Laboratory (CSAIL). Prior to joining MIT, Dr. Sontag was an Assistant Professor in Computer Science and Data Science at New York University from 2011 to 2016, and a postdoctoral researcher at Microsoft Research New England. Dr. Sontag received the Sprowls award for outstanding doctoral thesis in Computer Science at MIT in 2010, best paper awards at the conferences Empirical Methods in Natural Language Processing (EMNLP), Uncertainty in Artificial Intelligence (UAI), and Neural Information Processing Systems (NeurIPS), faculty awards from Google, Facebook, and Adobe, and a National Science Foundation Early Career Award. Dr. Sontag received a B.A. from the University of California, Berkeley.

Robert Truog, MD

Robert Truog is the Frances Glessner Lee Professor of Medical Ethics, Anaesthesiology & Pediatrics and Director of the Center for Bioethics at Harvard Medical School. Dr. Truog received his medical degree from the University of California, Los Angeles and is board certified in the practices of pediatrics, anesthesiology, and pediatric critical care medicine. He also holds a master's degree in philosophy from Brown University and an honorary master of arts from Harvard University.

As director of the Center for Bioethics, he has overall responsibility for the Center's many activities, including the master of bioethics graduate program, the bioethics fellowship program, required courses in medical ethics and professionalism for Harvard medical students, and the Center's many workshops, seminars, and public forums. As chair of Harvard University's Embryonic Stem Cell Research Oversight Committee (ESCRO), he is engaged in the interesting and difficult challenges of defining the ethical parameters of stem cell research and regenerative biology

Dr. Truog practices pediatric intensive care medicine at Boston Children's Hospital, where he has served for more than 30 years. He has published more than 300 articles in bioethics and related disciplines, and his writings on the subject of brain death have been translated into several languages. He authored current national guidelines for providing end-of-life care in the intensive care unit. He was principal investigator on the recently completed NIH study Toward Optimal Palliative Care in the PICU Setting. His books include "Talking with Patients and Families about Medical Error: A Guide for Education and Practice" (2010, JHUP, translated into Italian and Japanese), and "Death, Dying, and Organ Transplantation" (2012, Oxford).

He lectures widely nationally and internationally, and is an active member of numerous committees and advisory boards. He has received several awards over the years, including the the William G. Bartholome Award from the American Academy of Pediatrics, the Christopher Grenvik Memorial Award, and the Shubin-Weil Master Clinician-Teacher Award, both from the Society of Critical Care Medicine. In 2013 he was honored with the Spinoza Chair at the University of Amsterdam.

Mark Johannes van der Laan, PhD

Mark Johannes van der Laan, PhD, is the Jiann-Ping Hsu/Karl E. Peace Professor of Biostatistics and Statistics and Co-Director, Center for Targeted Machine Learning and Causal Inference at the University of California, Berkeley. He has made contributions to survival analysis, semiparametric statistics, multiple testing, and causal inference. He also developed the targeted maximum likelihood methodology. He is a founding editor of the Journal of Causal Inference. He received his Ph.D. from Utrecht University with a dissertation titled "Efficient and Inefficient Estimation in Semiparametric Models". He received the COPSS Presidents' Award in 2005, the Mortimer Spiegelman Award in 2004, and the van Dantzig Award in 2005.

Robert Vandersluis

Robert Vandersluis is VP of Artificial Intelligence Ethics and Policy at GSK, where he researches ethical and public policy issues surrounding the development and deployment of AI system in the pharmaceutical sector. Building on the success of the GSK.ai Fellows Programme, Mr. Vandersluis is also in the process of launching post-doctoral AI Ethics and Safety Fellowships at Stanford University. The goal of these fellowships will be to undertake original research to help GSK and the rest of the pharma industry to deliver AI-based interventions that are inclusive, empowering, and safe.

Prior to his work in AI, Mr. Vandersluis managed a £20 billion investment portfolio at GSK and a £20 billion balance sheet at FCE Bank PLC, where he was Treasurer. He has also served on the board and chaired the Audit & Risk Committee of LPPI, which has £22 billion of assets under management. Within the charitable sector, Mr. Vandersluis served on the boards and investment committees of the Pensions Trust, which has £13 billion under management, and the Walcot Foundation, which has a £120 million endowment.

Mr. Vandersluis is currently undertaking studies in Practical Ethics at the University of Oxford, where he focuses on issues related to artificial intelligence. Previously, Robert studied economics, politics, philosophy, and public policy at the London School of Economics, the University of Michigan, the University of Cambridge, and Harvard University's Kennedy School of Government.

Li Wang, PhD

Li Wang is a Senior Director and the Head of Statistical Innovation group in AbbVie. Dr. Wang is leading Design Advisory which provides strategic and quantitative consulting as requested to all Development teams in all Therapeutic Areas to facilitate innovative thinking and complex innovative design evaluation. He also co-leads Development Advanced Analytics capability in AbbVie to drive Machine Learning and Advanced Analytics research and application in Development. Prior to this senior leadership role, Dr. Wang led Immunology and Solid Tumor statistical design and strategy discussions and multiple ML, RWE and Bayesian innovation projects from 2017 to 2019. From 2006 to 2017, he contributed to and subsequently led several NDAs and SNDAs including blockbusters Eliquis, Onglyza and Rinvoq. Dr. Wang is enthusiastic in teaching statistical courses to non-statisticians, and investigating/ promoting novel statistical and machine learning methodologies. He is also active in statistical communities as chair of ICSA Midwest Chapter and executive committee for DIA VJC.

Marinka Zitnik, PhD

Marinka Zitnik is an Assistant Professor at Harvard University with appointments in the Department of Biomedical Informatics, Broad Institute of MIT and Harvard, and Harvard Data Science. Dr. Zitnik's group investigates machine learning with a current focus on creating foundational models that require infusing structure and domain knowledge. This research develops artificial intelligence tools to guide therapeutics discovery and creates new avenues for giving the right patient the right treatment at the right time to have medicinal effects consistent from person to person and with results in the laboratory. Her research won best paper and research awards from the International Society for Computational Biology, International Conference on Machine Learning, Bayer Early Excellence in Science Award, Amazon Faculty Research Award, Roche Alliance with Distinguished Scientists Award, Rising Star Award in Electrical Engineering and Computer Science, and Next Generation in Biomedicine Recognition, being the only young scientist who received such recognition in both EECS and Biomedicine. Learn more about Dr. Zitnik's lab: https://zitniklab.hms.harvard.edu

KEYNOTE ADDRESS

RESPONSIVE REGULATIONS IN THE AGE OF EXPONENTIAL TECHNOLOGIES

M. Khair ElZarrad, Center for Drug Evaluation and Research (CDER), U.S. FDA

Abstract. The talk will provide an overview of the current and potential uses of technological innovations, with a particular focus on artificial intelligence (AI) across drug development. Current FDA activities to facilitate the responsible use of AI, as well as perspectives on the management and regulation of innovative technologies will be discussed. The potential utility of AI to support the use of real-world evidence, to facilitate more efficient clinical trials, and to advance digital health technologies will be described. The talk will also outline policy considerations to help establish a robust regulatory ecosystem that not only manages innovations, but also encourages exploring the full potential of such advancements.

PLENARY SESSION I

DATA SCIENCE AND COMMON DISEASES

Anthony Philippakis, Broad Institute

Abstract. While the 21st century has been the golden age of drug development for rare diseases and cancer, there has been markedly less progress in developing new therapies for common diseases. A significant reason for this is the high cost of common disease clinical trials, which often require tens of thousands of patients. In this talk, we will review recent efforts at creating instruments based on genomic and clinical data that better enable selection of patients that both have higher rates of events, as well as a greater response to therapies. We will also review efforts at creating the Terra data platform which enables researchers to store, share and analyze genomic and clinical data.

ML AND AI IN CLINICAL DEVELOPMENT AND STATISTICAL INNOVATION

Li Wang, AbbVie

Abstract. ML and AI are widely used in technology industry now and are finding their ways into drug development in pharmaceutical industry especially in manufacturing and discovery. In clinical development, how and where to appropriately apply ML/AI and what value it can bring to the business remain big questions without clear answers to researchers.
For traditional clinical statisticians, it is not only a big challenge but also an exciting opportunity. AbbVie's statistical innovation group is revamped with the new vision and focus on Digital, RWD, ML and Bayesian methodologies. We are closely collaborating with different Development functions to leverage multiple data sources to help make smarter decisions. Experiences and some use cases will be shared.

PRACTICAL APPROACHES FOR PREPROCESSING COMPLEX AND HIGH DIMENSIONAL HEALTHCARE DATA IN PREDICTIVE MODELING USING MACHINE LEARNING TECHNIQUES, WITH APPLICATIONS TO COVID 19 MORTALITY AND WEARABLE DEVICE DATA

Javier Cabrera, Rutgers University

Abstract. Clinical and healthcare data that are essential to generate valuable medical information based on deep learning and other machine learning methods are typically complex, high dimensional and often noisy. The sources of such data may be mobile devices, registries or even social media. While there exist many technology solutions to handle the retrieval of the data, it may require substantial effort to prepare the data as input for the machine learning algorithms. In this talk, we intend

to focus on novel approaches to healthcare data preprocessing, with special reference to data collected from wearable devices. Specific questions addressed include choice of scale, temporal aggregation, and optimal data transformation. Further illustrations will be provided on the issues surrounding the preprocessing of data collected for use in predicting Covid-19 death outcomes. It is noted that proper preprocessing of complex data maybe be a critical factor for the successful implementation of AI solutions in medical research.

AI IN HEALTH: FROM DIGITAL HEALTH TO TRANSFORMING LIFE SCIENCES – DATA AND METHOD CONSIDERATIONS

Usama Fayyad, Northeastern University

Abstract. Health in general, including public health, healthcare delivery and administration, wellness in various areas, and Life Sciences have been experiencing digitization of processes, tracking, and data for analysis. While such digital transformations are often lagging behind other industries, practitioners in all these areas of Health are suffering from the typical after effects of digitization: much larger volumes of data being generated, a strong desire to leverage AI and algorithms to help deal with these data and use it to advance practices, and a dearth of capabilities, plans, strategies for how to do this. This results in data assets that deliver little to no value to the fields and in fact turning data from assets to liabilities and cost drivers. In this talk we outline a wide set of opportunities and examples of how they can be addressed. Taking a more systematic, standardized, and pragmatic approach to how we effectively enable the transition to more computationally intensive and AI-enabled approaches is likely one of the grand challenges for truly transforming Health and moving it from the manual/ artisanal practice to a digitized, data-enabled, and AI-aided approaches. In addition to some case studies, we discuss the risks and dangers that need to be addressed in such efforts.

PLENARY SESSION II

FINDING NATURAL EXPERIMENTS IN OBSERVATIONAL DATA

Adler Perotte, Columbia University

Abstract. Evidence on clinical treatments based on real world data has the great benefit of representing effectiveness in a broader population than is typically studied in prospective experiments. However, confounders that influence both the choice of treatment and the outcomes of interest can disguise the truth about a given treatment's effectiveness. Methods such as propensity score matching and weighting offer approaches to minimize the effect of these confounders, but leave open the question of which populations of people should be compared. In this work, we explore a computational approach to identifying the subgroup from all treatment arms of an observational study that are comparable, effectively finding the natural experiment for the target treatments. This work leverages a deep generative model based on generative adversarial networks and results in the identification of a comparison distribution that minimizes the variance of estimates such as the average treatment effect.

FRAMEWORK FOR ASSESSING THE REPRODUCIBILITY OF OBSERVATIONAL COMPARATIVE EFFECTIVENESS RESEARCH

Asieh Golozar, Odysseus Data Services, Inc

Abstract. The value of observational research lies in its ability to support patient care by producing clear signals regarding which treatments work for which patients. A federated network allows us to investigate if the signal generated from one site is generalizable across the multiple sites. The realization of this value relies on differences in source data, design, and the operational and analytic considerations. The factors that determine the impact of these considerations on the reproducibility of research findings are not well understood. A greater understanding of factors affecting reproducibility research will increase recognition of the value of large-scale healthcare databases for addressing a wide range of health-related questions. The reproducibility framework will allow decision makers to better evaluate how much trust should be placed in observational

research results. Implementation through validated software tools will allow scalable approach to assess reproducibility at a large scale. The success of the application of this framework will enable advances in theories about the factors affecting reproducible observational research.

The talk will focus on the following:

- The framework development process
- Measurement/quantification of the degree of heterogeneity in research findings attributed to data, design, operationalization, and analytics considerations

TOWARDS INTEGRATION OF TARGETED LEARNING AND CAUSAL INFERENCE IN DRUG APPROVAL PROCESS AND SAFETY ANALYSIS

Mark J. van der Laan, University of California, Berkeley

Abstract. Targeted Learning represents a general roadmap for accurately translating the real world into a formal statistical estimation problem, and a corresponding template for construction of optimal machine learning based estimators of any desired target causal estimand combined with formal statistical inference. It is flexible by being able to incorporate high dimensional and diverse data sources. To optimize finite sample performance, it can be tailored towards the precise experiment and statistical estimation problem in question, while being theoretically grounded, optimal, and benchmarked. We provide a motivation, explanation, and overview of targeted learning; the key role of super-learning; and discuss SAP construction based on targeted learning. We also discuss a Sentinel and FDA RWE demonstration project of targeted learning.

PLENARY SESSION III

AUGMENTING DRUG DEVELOPMENT USING CLINICAL TRIAL DATA MINING AND MACHINE INTELLIGENCE

Shameer Khader, AstraZeneca

Abstract. Clinical trial data represents an important yet often less valued analytics strategy in biomedical data science. We have been exploring the clinical trial big data compendiums, integrating data siloes, and utilizing machine intelligence to aid clinical development. We have combined data from multiple sources: including open clinical trial catalogues (ClinicalTrials.Gov, Aggregated Analytics of Clinical Trials.gov, etc.) and real-world data to gain novel insights and develop algorithms using modern machine intelligence approaches. We have developed a range of AI-driven methods to accelerate different facets of drug development and clinical trials. Some of our successful examples include clinical intelligence-driven patient engagement, understanding the complex landscape of early-to-late endpoints in oncology and developing scalable computing solutions to mine clinical trial big data (TrialGraph, ClinicalTrials2Vec). To conclude, clinical trial data mining is critical for improving future clinical trials across different therapeutic areas. Drug development teams could use the clinical intelligence insights and models to augment trials, improve patient engagement, reduce side effects, and accelerate drug discovery.

AI-DRIVEN CLINICAL DOCUMENTATION WITH MEDKNOWTS FOR CHEAP AND SCALABLE COMPARATIVE EFFECTIVENESS AND SAFETY MONITORING STUDIES

David Sontag, MIT

Abstract. A key challenge for scaling up the generation of real-world evidence from electronic medical records is deriving high-quality outcomes, interventions, and confounding factors from noisy and unstructured clinical data. This is needed for both comparative effectiveness and safety monitoring studies. A common approach is to retrospectively derive the requisite structured data using algorithms for electronic phenotyping and natural language processing. Can we instead collect the data we need *prospectively*, during routine clinical care, while being unobtrusive, saving clinicians' time, and fitting into existing clinical workflows? I describe a new clinical documentation system, MedKnowts, that aims to do precisely this. MedKnowts is an integrated note-taking editor and information retrieval system which automatically captures structured data while still

allowing users the flexibility of natural language. MedKnowts leverages this structure to enable easier parsing of long notes, auto-populate text, and perform proactive information retrieval. I'll describe our findings from an initial pilot at the Emergency Department of Beth Israel Deaconess Medical Center, where MedKnowts was used to write the notes for over 1500 patients, and our current work on re-envisioning MedKnowts for oncology.

THE EFFICACY ARM IN SILICO: OPPORTUNITIES AND CHALLENGES WITH USING REAL-WORLD DATA IN DRUG DEVELOPMENT AND COMPARATIVE EFFECTIVENESS

Ravi B. Parikh, University of Pennsylvania

Abstract. Clinical trial optimization is an unmet need in oncology: potentially revolutionary therapeutics take years to impact patient lives and drug development budgets are skyrocketing. Artificial intelligence methods based on real-world, clinico-genomic data are being increasingly used to accelerate the drug discovery process. However, issues with underlying data generation processes and unfairness in algorithm outputs exist. In this talk, I provide examples of promising use cases of real-world data and AI methods in drug development and comparative effectiveness, highlight underlying biases in real-world data, and propose best practices moving forward.