Session abstract

Session: Data science research ethics

September 20, 17:30-19:00

New challenges arise in research as data sets have grown in size and complexity. Data from different sources and public data are used to explore research questions. Are you prepared to address the emerging ethical issues surrounding research with big data? The research process should include a pre-specified study design and analysis plan and follow a systematic approach, including careful documentation for reproducibility, with data protection measures in place. Collaborations across multiple perspectives (scientific, statistical, computational, statistical, ethical) are needed. Training programs on big data ethics for graduate students should be offered and more awareness of data ethics challenges is needed for researchers. This session aims to provide an overview of some of these challenges.

Organizer: Marianne Huebner, Department of Statistics and Probability, Michigan State University, USA

Invited speakers:

Managing research data for transparency and reusability
Scout Calvert, University Library, Michigan State University, USA

Said to promote reproducibility and prevent fraud, data sharing is becoming a scientific norm and an expectation from funding agencies. There's also evidence it can help the careers of researchers. But sharing data is not so easy as simply sharing files. This presentation will provide some strategies for managing data so it can be ethically shared, understood, and reused.

Good data science practice: Moving towards a code of practice for drug development
Mark Baillie, Novartis, Switzerland
mark.baillie@novartis.com

There is growing interest in data science and the challenges that could be solved through its application. The growing interest is in part due to the promise of "extracting value from data". The pharmaceutical industry is no different in this regard reflected by the advancement and excitement surrounding data science. Data science brings new perspectives, new methods, new skill sets and the wider

use of new data modalities. For example, there is a belief that extracting value from data integrated from multiple sources and modalities using advances in statistics, machine learning, informatics and computation can answer fundamental questions. These questions span a variety of themes including: disease understanding (i.e. "precision" medicine, disease endo/phenotyping, etc.), drug discovery (i.e. new targets and therapies), measurement (i.e. multi-omics, digital biomarkers, software as a medical device, etc.), and drug development (i.e. dose-exposure-response, efficacy, safety, compliance, etc.). By answering these fundamental questions, we can not only increase knowledge and understanding but more importantly inform decision making; accelerating drug and medical device development through data-driven prioritisation, precise measurement, optimised trial design and operational excellence. However, with the promise of data science, there are also several obstacles to overcome, especially if data science is to live up to this promise and deliver a positive impact. These obstacles include consensus on a common understanding of the very definition of data science, the relationship between data science and existing fields such as statistics and computing science, what should be involved in the day to day practices of data science, and what is "good" practice. The talk will explore these issues with the aim of opening a dialogue on good data science practice.

Data Science Research Ethics and the Challenges of Inference, Public Data and Consent

Jacob Metcalf, Data & Society Research Institute

Data science, and the related disciplines of machine learning and artificial intelligence, are founded on the assumed availability of massive amounts of data. The scientific and economic justification for collecting and using all that data is deceptively simple: we can infer expensive- and hard-to-know data from cheap- and easy-to-know data and make predictions and automated decisions on the basis of the patterns we find. When that data is about human behavior, that inferential step is ethically fraught because it often involves data that is ubiquitous (social media, geolocation, biometrics, etc.) being used to predict traits that are from an entirely different context (race, religion, sexual preference, gender, etc.), and typically without knowledge or consent. This is a highly complex ethical challenge, yet our research ethics norms and regulations were written for a different paradigm of scientific research. In this talk, I will illustrate this dynamic with several cases of data science research ethics controversies and consider how we might establish new practices for ethical research.