PhUSE US Connect 2020

Paper AS13

Demystifying the Data World of Biobanks with Machine Learning

Himanshu Pandya, Biogen, RTP, USA
Bob Engle, Biogen, Cambridge, USA

ABSTRACT

Since last few decades Biobanks have become an important resource in medical research supporting many types of contemporary research like genomics and personalized medicine. Biobanks give researchers access to data representing a large population. They collect, catalogue and store biological samples and associated clinical data.
One such Biobank is the UK Biobank (UKBB). UKBB is one of the 10 largest Biobanks of the world and with around 500K subjects RWD. Its data is currently being used by many researchers (including Biogen) in the life science industry.

This paper illustrates overview of the UKBB, the associated data curation process, salient features of the process, challenges faced and how ML is being leveraged to gain invaluable data insights.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the authors. The interpretations of standards and procedures contained in this paper are those of the authors. Any views and recommendations stated within this paper are those of the authors, and they do not represent the positions of their employer.

INTRODUCTION

Biobanks are type of biorepositories that stores biological samples (usually human) for use in research. They have become an important resource in medical research, supporting many types of contemporary research like genomics and personalized medicine. This paper presents what Biobanks are and their importance, along with a brief overview on UK Biobank (UKBB), associated data curation process and challenges faced during the process. Further we discuss how ML is being leveraged to gain invaluable insights on data such as - patient population, disease history, progressions etc.

This paper is broken into the following sections:

- What are Biobanks and significance of Biobanks in the life science industry.
- A glimpse of UKBB and UKBB data.
- Biogen's data harmonization and curation process - from analysis and regulatory submission perspective
- Data usage at Biogen and associated ML uses cases.
- Conclusions

WHAT ARE BIOBANKS?

- Biobanks are collections of human biological tissue specimens and related health data – BIG data.
- In addition to biological samples, appropriate sample and donor-related information is also collected with the donor's permission.
- Information has been collected, for example, by questionnaires, during sampling, in the context of a medical examination, or during hospitalization.
- The data is stored in accordance with proper data processing and management practices, as defined by law.
- Biobanks act as data stewards, and are responsible for ensuring patient privacy and ensuring that samples, or biospecimens, are only used for intended purposes
- Different biobanks collect different types of samples and information. The types of information and samples collected depend on the specific purpose of the biobank. For example, some biobanks are specific to a disease, such as cancer. Other biobanks are population based and contain samples and information from people in a specific population or region

IMPORTANCE OF BIOBANKS

- Biobanks are a way in which ordinary people, as well as patients, can become involved with medical research. By donating biological samples and specimens, participants can help accelerate research studies and find treatments for diseases
- Biobanks serves as libraries for researchers. Therefore, the time and resources needed to recruit new participants for each research study is greatly reduced because samples and corresponding medical information are available in one place. By making sample collection and patient recruitment more efficient, better studies can be performed in a timely fashion.
- Biobank research hopes to provide novel insights into the genetic component of disease, ultimately leading to a more personalized approach to healthcare
- Samples in biobanks and the data derived from those samples can be used by multiple researchers for cross purpose research studies.
- The structure of biobanks fosters cross-collaboration between disease advocacy organizations and research scientists. Biobanks produce a synergy that hastens the research process. Making treatments or cures to genetic conditions attainable in the future.
- Biobanks collect, catalogue and store biological samples and associated clinical data. The specimens and data are collected from patients who are part of epidemiology studies or clinical trials, individual laboratory research studies, or who have agreed to be part of a patient cohort.
- Data helps researchers to gain a molecular understanding of the stages of disease progression. It helps researchers to develop biomarkers for diagnostic and prognostic purposes and to identify novel therapeutic targets for drug discovery.
- Biobanks are of great significance to the genetic counseling community as this type of research has the potential to improve diagnostic and treatment options for a broad range of genetic diseases.

UK BIOBANK (UKBB) – AN OVERVIEW

- UK Biobank is a prospective cohort study with deep genetic and phenotypic data collected on approximately 500K individuals from across the United Kingdom, aged between 40 and 69 at recruitment.
- Volunteers are being followed for years to record their demographic, medical, and other health-related information. This is a longitudinal study with some parts of the data still being gathered. Because of the large number of participants and the long follow-up time, managing and interpreting the data in the archive presents many challenges related to the data size, format complexity, feature heterogeneity, and sampling incongruence of the observations
- A rich variety of phenotypic and health-related information is available on each participant, including biological measurements, lifestyle indicators, biomarkers in blood and urine, and imaging of the body and brain. Follow-up information is provided by linking health and medical records. Genome-wide genotype data have been collected on all participants, providing many opportunities for the discovery of new genetic associations and the genetic bases of complex traits.
- It offers incredible research opportunities for the entire worldwide scientific community. The major objective of the UK Biobank is to improve the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses, such as cancer, heart disease, diabetes, depression, etch
- UK Biobank data will allow researchers from around the world to conduct research that leads to better strategies for the prevention, diagnosis and treatment of a wide range of life-threatening and disabling conditions

TYPES OF DATA COLLECTED

- Population Characteristics
- Sociodemographic
- Lifestyle and environment
- Family history
- Psychosocial factors
- Health and Medical History
- Physical Measures
- Cognitive Functions
- Imaging
- Biological Samples
- Procedural Metrics
- Genomics
- Online Follow-up
- Health-related outcomes

DATA HARMONIZATION AND CURATION PROCESS

With the growing UKBB data popularity, there was a huge demand among wide range of teams/stakeholders across Biogen to use the UKBB data for various analyses. Which created an urgent need to develop a central data repository which would cater data needs to all cross-functional teams within Biogen. The importance of the central data repository was to ensure a single 'gold standard' copy of data that would be the basis for all analyses. This helps to ensure traceability to the original source and reassures analysts that the data is correct.

As a result, project "UKBB Data Curation" originated.

The vision was to create an innovative platform with inbuilt centralized RWD (Real World Data) repository/warehouse. The platform would be scalable, auditable, reproducible and software agnostic. The platform would also have the potential of integrating future RCT (Randomized Clinical Trials) studies for regulatory submissions.

Some of the salient features

- A complete life cycle of data process, wherein the entire process is reproducible, traceable, auditable and scalable.
- Central data hub/repository, which is designed and developed based on expertise/need/input from Researchers/Scientists/Biostatisticians
- Automated process based on HPC (High Performance Computing) environment, Grid (SAS ®) platform and Biogen's indigenous tools, to generate data files (SAS7BDAT, CSV) with concise variable/column names and labels from the raw/source data files.
- Software agnostic (SAS ®/R/Python) and platform agnostic (HPC/Grid/AWS/Data Bricks) to cater wide range of end users and their analysis requirements.
- Have the potential of integrating the data with future RCT studies for regulatory submissions and provides flexibility of creating CDISC/SDTM compliant data (per business needs) for regulatory agencies submission purpose.
- Ability to act as a bridge between RCT/RWE programs and support the agencies framework for evaluating the potential use of RWE/RWD in drug approval processes

Data Usage and ML Implementation

Based on the curated data, analysis datasets were created by integrating broad range of variables of interests to enable the ability to answer novel research questions.

Analysis datasets representing two categories – Phenotype, Genotype were developed. The structure of the datasets was close to the regulatory agency submission standards:

- **Phenotype analysis datasets**
  Integrated datasets representing data endpoints from wide range of data such as -
  - Socio Demographics
  - Patient Characteristics
  - Physical Measures
  - Medical History
  - Disease History
  - Cognitive Test data such as Fluid intelligence, Reaction time, Numeric memory, Prospective memory, Symbol digit substitution, Trail making
  - Online Cognitive Test data – Fluid intelligence, Numeric memory, Symbol digit substitution, Trail making

- **Genotype analysis datasets**
  Based on all collected Genotype data and selected Socio Demographics data endpoints

Based on different research requests, additional data operations were performed

- Recode all data entries to a descriptive form (for example 0/1 codes to female/male)
- Check ranges for each variable
- Recode ethnic groupings
- Recode education variable to years of education
- Create “Time to Event” variables

ML APPLICATION AND IMPLEMENTATION

Supervised and unsupervised ML learning concepts were applied for this project.

- **Unsupervised Learning**
  Basic goal was to discover unknown new sub-groups among the data based on certain disease characteristics i.e. different components of cognitive tests. This was efficiently achieved by means of “Clustering” **K-Means**. Clustering allowed us to identify which observations are alike, and potentially categorize them.
  Cluster based on Phenotype data - age/education/baseline cognitive results were generated for better understanding of the underling subgroups.

- **Supervised Learning**
  Under the supervised learnings, Linear Regression was used to understand relationship among the variables of interests. Dependent variables of interest were Cognition decline progression variables. Predictors consisted of various baseline Cognitive test variables and Socio Demographic variables such as - age, education, sex etc.

CONCLUSIONS

- Biobanking is changing the world by influencing our knowledge about human health, disease, drugs, personalized medicine, and more. Biobanks have been heralded as essential tools for translating biomedical research into practice, driving precision medicine to improve pathways for global healthcare treatment and services. Biobanks have humongous data size, it's not unusual to see data size of multiple terabytes and can have very complex data structure.

- It is a big challenge to harmonize the data elements and extract useful information from such complex data. Due to the massive size of the data, large number of participants and the long follow-up time, managing and interpreting the data presents many challenges, especially during the data curation process. Challenges can be related to the data size, data structural, format complexity, variable naming conventions etc. Also, the large number of observations and the complex composition of the data elements make it difficult for researchers to statistically mine and computationally generate precise, reliable and consistent inferences.

- "Bigger the challenge, the bigger the opportunity".
  Biobank data is goldmine for researchers/scientists and can be foundation for many important discoveries. Rewards and benefits are huge and can be life changing for human race. With the advancement in technology i.e. AI/ML, statistical software i.e. R/Python/SAS®, robust cloud based computational processing power and availability of excellent data mining tools it's now **HIGHLY** possible to overcome the known/unknown challenges to achieve the expected results.

ACKNOWLEDGMENTS
We are very thankful to the UKBB website **https://www.ukbiobank.ac.uk/** for content sharing.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Authors can be contacted at:

Himanshu Pandya, Biogen, RTP, USA
Himanshu.Pandya@Biogen.com

Bob Engle, Biogen, Cambridge, USA
Bob.Engle@Biogen.com