COVID-19 Evidence Accelerator Collaborative

Lab Meeting # 24

Thursday, December 3, 2020, 3 - 4:00 pm ET

Call Summary

Lab Meeting 24 Overview

The December 3rd therapeutics lab meeting featured a deep dive into Artificial Intelligence (AI). First, Ben Birnbaum of Flatiron Health provided an overview of some basic concepts and applications of AI & Machine Learning (ML). Next, Cindy Wang and Samson Mataraso presented on how Dascena utilizes the power of ML algorithms to respond to COVID-19, including developing an algorithm for identifying COVID-19 patients needing intubation within 24 hours which was recently granted Emergency Use Authorization (EUA). Finally, Ofer Mendelevitch of Syntegra described their use of AI for generating synthetic data models using real-world data.

Machine Learning (ML) in Real-World Evidence (RWE)
Ben Birnbaum, Flatiron Health

Introduction to Artificial Intelligence & Machine Learning terms

- **Artificial Intelligence (AI)** – Computer systems able to perform tasks that normally require human intelligence.
- **Machine Learning (ML)** – Use of statistical techniques to give computers the ability to learn without being explicitly programmed.
- **Deep Learning** – A set of ML techniques using multi-layer neural networks.
- **Classification** – A common type of ML (especially useful in RWE) which predict a label form a discrete set of classes for a given example.
    - **RWE Example –** Has a patient received hydroxychloroquine? Has a patient had COIVD-19 symptoms for longer than 12-weeks?
- **Supervised Approach –** Approach to ML that involves training data (e.g., identifying and assigning labels to spam emails) & feeding the trained data into a machine learning algorithm which uses statistical methods (e.g., logistic regression, decision trees, neural networks) to predict labels, and ultimately create a trained model.

Flatiron's Utilization of AI/ML in RWD/RWE

- Flatiron's Perspective of ML is that it can be a powerful tool for increasing the impact of RWE as long as we recognize that it can introduce additional risk that needs to be understood & mitigated.
- ML can be used to automate the extraction of information from large unstructured datasets (e.g., identifying/extracting metastatic patients from EHR data). Automating these processes is less time consuming than manual/human processing of these data.

Case Study: Machine Learning for Cohort Selection

- Prior to using ML for cohort selection, Flatiron used structured codes to identify patients with breast cancer, and clinical extractors used unstructured data to confirm metastatic disease.
- Using a supervised approach, Flatiron used ML to identify metastatic breast cancer patients from unstructured EHR data.
    - The model was trained using short phrases found in EHR that were related to patients' metastatic diagnosis such as "stage iv breast ca", "proven metastatic", "bone mets", and "no evidence of metastatic disease".
    - Phrases were assigned values (1 for metastatic and 0 for non-metastatic) and logistic regression was performed.
    - The model removed patients it did not predict to have metastatic disease & clinical extractors confirmed those the model predicted should be included in the cohort.
    - Various bias analyses/mitigation steps were (and continue to be) applied to ensure the cohort selection process remains unbiased.
- ML successfully made the cohort selection process more efficient & there was a low rate of false negative identification.

Machine Learning Algorithms for COVID-19

Cindy Wang & Samson Mataraso, Dascena

Overview of Dascena's Use of ML

- Dascena is a medical technology company that creates & applies ML algorithms to clinical data to diagnose complex conditions → improved outcomes by providing earlier & more accurate diagnosis.
- Using ML to respond to COVID-19:
    - Dascena created & validated algorithms to predict treatment benefit for COVID patients, mortality prediction, ventilator use, etc.
    - Emergency Use Authorization (EUA) issued to Dascena's ML algorithm designed to predict a patient's need for ventilator use within 24 hours (COViage).

Dascena's Approach to Algorithm Development

1. Indication specific dataset construction
    a. Gathering & merging data from various databases (e.g., hospitals, cancer centers, STD clinics, GP visits) that are determined by the defined unmet need or problem, setting in which the solution will be deployed, and patient population for which the algorithm is being designed.

2. Dataset labeling & characterization
    a. Data labeled based on gold standard (ICD codes, clinical criteria, etc.).
    b. Features for development are extracted from the raw data.
    c. Numerical processing such as standardization, normalization, etc. are used to refine the features.
    d. Quality checks performed.
3. Algorithm generation & testing
    a. Algorithm generated & refined using supervised ML methods.
    b. Features are refined & best parameters for features are chosen to ensure optimal performance is maintained at clinically acceptable levels.
    c. Operating point chosen to balance sensitivity & specificity.
    d. Algorithm is tested on a test dataset to ensure it will perform as it is intended to on unseen RWD & prevent overfitting.

COViage Algorithm

- ML algorithm developed by Dascena issued an EUA. The algorithm predicts hemodynamic instability or respiratory decompensation in patients with COVID-19.
- Developed from a population of patients hospitalized with COVID-19.
- Vital signs and demographic data were abstracted from EMR data & converted into features of the algorithm.
- Algorithm was developed using a supervised ML approach.

Synthetic Data Generation Using Advanced Deep Generative Models

Ofer Mendelevitch, Syntegra

Modern Language Modeling & Synthetic Data Generation

- New solution to create synthetic derivatives of RWD that is privacy preserving, with validated precision, and explainable.
- **Modern Language Models** – language models like GPT and BERT are unsupervised algorithms from the field of natural language processing, that are used to predict the next word in a sentence by assigning probabilities to possible proceeding words.
    - These algorithms can be used to predict patient outcomes by turning medical data into "patient sentences" and assigning probabilities to "words" which are health outcomes.
- Language models work very well in practice as they can handle both big datasets & small cohorts, support any type of variable (numeric, missing variables, etc.), and allow for greater portability/privacy preservation due to generation by random sampling.

Synthetic Data Case Study: The Effect of Digoxin on Mortality in Patients with Heart Failure

- Using synthetic & real data points to model 6800 patient records with 71 variables.
- Dimensionality reduction with UMAP – Takes medical record information and turns it into 3-dimensional visual using real and synthetic data points; this is a great visual aid to help understand synthetic data fidelity and privacy and especially in rare cohorts.

- To ensure quality of the synthetic data the outcomes (i.e., mortality or worsening heart failure) from the synthetic dataset are compared to outcomes from real data using Pairwise Correlation, Kaplan Meier Survival Analysis, & Predictive Modeling Analysis (gradient boosted trees).

National COVID Cohort Collaboration (N3C) Dataset of COVID Full EMR

- Working with NIH to create a synthetic dataset from EHR data.

Currently includes 2.1 million patient records, 292,000+ COVID patients, more than 2 billion rows of data