Principal Investigator

First Name: Kin Wah
Last Name: Fung
Degree: MD
Primary Affiliation: National Library of Medicine
E-mail: craig.mayer2@outlook.com
Phone number: 9542705971
Address: 8600 Rockville Pike

City: Bethesda
State or Province: MD
Zip or Postal Code: 20894
Country: United States

General Information

Key Personnel (in addition to PI):
 First Name: Nick
Last name: Williams
Degree: Ph. D
Primary Affiliation: Nattional Library of Medicine, NIH

 First Name: Craig
Last name: Mayer
Degree: MS
Primary Affiliation: Nattional Library of Medicine, NIH

Are external grants or funds being used to support this research?: External grants or funds are being used to support this research.
Project Funding Source: NIH Office of AIDS Research
How did you learn about the YODA Project?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/system/files/coi_cm.pdf
https://yoda.yale.edu/system/files/coi_kwf.pdf
https://yoda.yale.edu/system/files/coi_nw.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT00540449 - TMC278-TIDP6-C209 - A Phase III, Randomized, Double-blind Trial of TMC278 25 mg q.d. Versus Efavirenz 600mg q.d. in Combination With a Fixed Background Regimen Consisting of Tenofovir Disoproxil Fumarate and Emtricitabine in Antiretroviral-naive HIV-1 Infected Subjects
2. NCT00543725 - TMC278-TIDP6-C215 - A Phase III, Randomized, Double-blind Trial of TMC278 25mg q.d. Versus Efavirenz 600mg q.d. in Combination With a Background Regimen Containing 2

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

Identification of Research Common Data Elements in HIV/AIDS

Narrative Summary:

This research looks at reusability capabilities of HIV research by reviewing trial information and how that information is collected, analyzed, and reported and how that trend has changed. The study compares that information found in trials with that in everyday medicine. The goal of this study is to look at the trials and everyday information and understand what information is reported in each and compare them to find commonalities and differences. This analysis should allow for the development of uniformity in reporting information that can be implemented to improve the combination of ideas and the use of increased knowledge to improve outcomes for HIV patients.

Scientific Abstract:

Background: Many efforts emerged that try to maximize the value of data collected during human clinical trials. Secondary analyses of multiple comparable trials can generate additional discoveries or lead to novel hypotheses. A key that allows data integration across studies is the development of Common Data Elements (CDEs) that are incorporated into study design and data collection. Objectives: This projects aims to support this modern trend by analyzing research CDEs applicable to the domain of HIV research. Study Design: This study will retrospectively analyze data elements collected during HIV studies and other real-world data sources relevant to the domain of HIV research. We will analyze data elements based on their appearance in data sources as well as the participant volumes, and other statistical and analytic values and range of values present in various data elements in order to see the applicability of a data element to medical and research. Participants: We will be analyzing all participant-level data and all data elements and data types from all studies. Main Outcome: The main objective is to recommend optimal data representation format that allows data scientist to easily integrate HIV research datasets across studies. The project builds on NLM efforts to be the epicenter for NIH data science. It utilizes existing NLM expertise in routine healthcare terminologies and clinical research informatics. Statistical Analysis: The analysis will include the identification of the counts of variables and how they are used in practice by looking at patient populations.

Brief Project Background and Statement of Project Significance:

Secondary analyses of individual trials and aggregated meta-analyses of multiple comparable trials can generate additional clinical discoveries or lead to novel hypotheses. In recent years, many efforts have emerged attempting to maximize the data-reuse value of data collected during human clinical trials. Modern trials now regularly include patient consent to generate de-identified patient-level datasets at trial completion and modern investigators make

this data available for secondary research use. HIV/AIDS interventional trials or observational studies are following this trend toward data re-use that unlocks new opportunities for data from completed studies. A key component that allows for data integration and reuse across studies is the development of research Common Data Elements (CDEs) that are incorporated into the original study design and data collection. This project aims to support this modern trend by documenting, cataloging and analyzing research CDEs applicable to the HIV/AIDS domain to better recommend CDEs to HIV investigators. By improving CDE use and quality among HIV investigators our research efforts can improve investigator initiated HIV research as well as retrospective data-reuse across HIV research use cases.

Specific Aims of the Project:

AIM 1: Evaluate usability of existing efforts that share de-identified patient-level HIV/AIDS human clinical studies (interventional trials or observational studies). Design research protocol to submit when requesting data from trial result platforms
Extract studies from existing trial results platforms (e.g., Immport) . Convert study format into common format
Analyze common data elements present in HIV/AIDS trials

AIM 2: Design clinical research informatics solutions that provide optimal syntactic format and semantic terminology bindings for sharing HIV/AIDS trial data
Compare formats used by trial result platforms and existing data standards (e.g., CDISC Study Data Tabulation Model, OMOP Common Data Model)
Evaluate overlap between research study data elements and EHR data and annotate HIV/AIDS data elements using SNOMED CT clinical terminology.

This project requires IPD as opposed to data dictionaries or case report forms as not only the presence of CDEs will be assessed, but also their significance. This is done by comparing summary information about respective common data elements and their usage in a study. We will do this by comparing values and volume of data associated with the CDE to assess differences that may be clinically relevant

What is the purpose of the analysis being proposed? Please select all that apply.
Summary-level data meta-analysis
Summary-level data meta-analysis pooling data from YODA Project with other additional data sources
Participant-level data meta-analysis
Participant-level data meta-analysis pooling data from YODA Project with other additional data sources
Develop or refine statistical methods

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

The criteria we looked for when selecting studies were completed or no longer active late phase HIV related trials with preference placed on intervention over observational studies. These datasets include Electronic Health Records (EHR) and claims data from CMS and the Greater Plains Collaborative, as well as the Lung HIV study from the National Heart Lung and Blood Insitutetrial platform, studies from the National Institute of Allergy and Infectious Diseases HIV Trial Networks, and multiple still to be determined clinical trials from Vivli and clinicalstudydatarequest.com.

Main Outcome Measure and how it will be categorized/defined for your study:

The study will look at clinical trial and EHR data to find the commonality in the data that is shared across different HIV/AIDs data sharing platforms. The analysis will identify common data elements (CDEs) by mapping synonymous data elements in multiple datasets and identify ing which data elements are present in multiple datasets. We will take those CDEs and analyze how the data collected for them is different by looking at differences in the values of responses, data frequency and data density for the CDE in different datasets. This will allow us to identify any uniformity and disconnects between the sources of HIV data. The study will use this analysis to develop a way to understand the significance of different data elements on a subset of accessible HIV studies. We will develop a set of comparative metrics based on the CDEs present in the set of studies and do a comparative analysis of those metrics, which may include things such as age, mortality, and comorbidities, etc., to

determine their relevance in drawing clinical observations from the data.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

The main predictors are the common and uncommon data elements. The independent variables will include the number and identity of the research common data elements being shared. This information will serve as key information for two major aims of the study as they will serve as key variables for understanding the linkages and differences between clinical trials and observational studies and how the information shared between the two is different and how they can be analyzed to understand a full uniform system of data variable.

Statistical Analysis Plan:

This research involves analysis of individual patient data in interventional and observational HIV/AIDs trials. Our focus is on larger trials for both treatments and vaccinations. Our project focuses on the differences in the sharing of information among observational studies and interventional trials and the data elements that are shared different among these studies as well as the common data elements that can be found in both. Also preference is given to later phase trials (phase 2 or 3) due to the requirements and likelihood of data sharing is vastly increased over phase 1 and early phase trials. We are interested in all types of patients and part of future analysis will include the stratification of patient type, but we are interested in currently including all patients.
One goal of the analysis plan is to describe the current state of the art in obtaining data from past HIV/AIDS studies by analyzing relevant trial results data sharing platforms. The second aim includes a detailed analysis of current (and future) data elements. This also allows for the analysis of data sharing platforms by understanding the data elements commonly and differently shared among different clinical trial and observational study platforms. The analysis plan includes the calculation of statistical counts of trial information sharing and the analysis of Common Data Elements throughout the interventional trial and observational study information. Further analysis includes secondary analyses of individual trials or aggregated meta-analyses of multiple comparable trials can generate additional clinical discoveries or lead to novel hypotheses. This project offers a comprehensive analysis of data elements found in clinical trials and observational studies that focuses on the analysis of present data compared to missing data and allows for an understanding of information sharing and develops a plan for the sharing of a uniform group of data elements. Our analysis will involve doing summary statistics of variables in an individual dataset and comparing that to similar variables in other datasets, to see the difference in results for different patient groups while also using a cluster analysis of actual variables to view similarities in variables collected. All analysis will be done by creating these summaries and overviews of datases individually and than compaing the results of each dataset by comparting the number and specifics of variables while also comparing max and min values and other summary statistics of the variables.
Our analysis will work by first looking across a multitude of data sources and identifying common data elements (CDEs) present across the HIV domain. A CDE is any data element present in multiple data sources. Once a set of CDEs are identified we will take individual patient records and do a variety of comparative analyses on the CDEs. This will include comparing a variety of descriptive statistics for actual item responses such as the maximum, minimum, and average age, counts of specific comorbidities, etc. We will also do a similar analysis on the metadata for these elements such as counts on frequencies and data densities of a specific element. Since this will be done for CDEs we will use similar metric analysis on multiple sources and allow for the comparison of the multiple trials, studies, and real world data. More advanced analytic strategies will also be used such as machine learning to further develop these comparisons. This is all in an effort to understand the value and amounts of data collected to showcase the usefulness of informatics in secondary research and dvelop tools and recommendations for HIV trial analysis.
Software Used:
RStudio
Project Timeline:

anticipated project start date: 7/1/19, analysis completion date: 8/1/19, date manuscript drafted and first submitted for publication: 9/1/20, and date results reported back to the YODA Project:9/1/20

Dissemination Plan:

We plan to report our results through presentations at medical conferences and publications in peer-reviewed medical journals. This includes potentialy the AMIA conference, the Journal of Advanced Medicine, the JOurnal of Current HIV Research, etc.

Bibliography:

1.Sheehan J, Hirschfeld S, Foster E, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016.
2.NLM. Common Data Element (CDE) Repository. 2014. http://cde.nlm.nih.gov (accessed Aug 12 2015).
3.Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). Journal of biomedical informatics 2015; 57: 88-99.
4.Huser V, Burke C, Nguyen M, Amos L. Annotation of Research Common Data Elements Using Clinical Terminologies. AMIA Annu Symp Proc 2017.
5.Huser V, Shmueli-Blumberg D. Data sharing platforms for de-identified data from human clinical trials. Clin Trials 2018: 1740774518769655.