

# Automated Generation of Core Outcome Measures for Clinical Trials

Anas ELGHAFARI<sup>a</sup> and Joseph FINKELSTEIN<sup>a,1</sup>

<sup>a</sup>*Icahn School of Medicine at Mount Sinai, New York, New York, USA*

**Abstract.** Common outcome sets are vital for ensuring usability of clinical trial results and enabling inter-study comparisons. The task of identifying clinical outcomes for a particular field is cumbersome and time-consuming. The aim of this work was to develop an automated pipeline for identifying common outcomes by analyzing outcomes from relevant trials reported at ClinicalTrials.gov and to assess the pipeline accuracy. We validated the output of our pipeline by comparing the outcomes it identified for acute coronary syndromes and coronary artery disease with the set of outcomes recommended for these conditions by a panel of experts in a widely cited report. We found that our pipeline identified the same or similar outcomes for 100% of the outcomes recommended in the experts' report. The coverage of the pipeline's results dropped only slightly (to 21 out of 23 outcome domains, 91%) when we restricted the pipeline to trials posted before the publication of the report, indicating a great potential for this pipeline to be used in aiding and informing the future development of core outcome measures in clinical trials.

**Keywords.** Clinical trials, common data elements, automation

## 1. Introduction

A major challenge in the design of clinical trials is the lack of consensus on which common outcomes should be reported. The importance of developing common data elements (CDEs), including common outcomes, has been emphasized by researchers in various fields of medicine [1-2]. A major obstacle to the development of common outcomes is the tremendous amount of time and effort required. The traditional approach involves assembling panels of subject matter experts who then embark on lengthy deliberations. This process does not typically include an automated large-scale analysis of trials relevant to the conditions for which the outcomes are being developed.

In this work, we seek to fill this gap by developing an automated pipeline for identifying outcomes using large-scale analysis of clinical trial outcomes from ClinicalTrials.gov. We evaluate the coverage of the pipeline by comparing the outcomes it identifies for "Acute Coronary Syndrome" and "Coronary Artery Disease" to a set of gold-standard outcomes proposed for these conditions by the American College of Cardiology Foundation (ACCF) and the American Heart Association (AHA) [3].

---

<sup>1</sup> Corresponding Author, Joseph Finkelstein, Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, New York 10029, USA; E-mail: Joseph.Finkelstein@mssm.edu.

## 2. Methods

### 2.1. *ClinicalTrials.gov Application Programming Interface (API)*

ClinicalTrials.Gov (CTG) is a repository of data from publicly and privately funded clinical trials. It contains records of 322,735 trials at the time of writing. CTG provides API that facilitates automated collection and easy parsing of trial data. We use two functionalities of the API. (1) The search functionality: allows the user to formulate a search request using a query term as either a “condition” or “other term”. This API returns a list of results that matches the query term. (2) The trial data downloading functionality: given a trial ID number (NCT number), this functionality allows the user to obtain the marked-up data of that trial in XML format (other formats are also available). This allows easy parsing of the primary and secondary outcomes, their definitions (if provided), and their time frames.

### 2.2. *Outcome Collection Pipeline*

The inputs to the pipeline is (a) a list of conditions (could also contain a single element), and (b) (optionally) a cutoff date for the trials that will be processed. For each condition an API call is generated to clinicalTrials.gov, a table of results is returned. Each row in the table includes, among other elements, the id number of a trial that matches the condition and the “first posted” date (the date on which the study appeared on CTG). The tables for each condition are aggregated into a single table while duplicates are removed. If a cutoff date is provided then the table is further filtered to include only trials whose “First Posted” date precedes the cutoff date.

The next step involves using the NCT number of each trial to generate an API request that returns the trial’s data in XML format. The xml is parsed to extract the primary and secondary outcomes. Outcome names are normalized with respect to capitalization, punctuation, and the inclusion of abbreviations. Finally, a mapping is constructed from each normalized outcome to a list of trials in which the outcome appears.

### 2.3. *Evaluation Method*

We evaluate the pipeline by comparing the outcomes identified by it to the set of outcomes proposed by the ACCF/AHA task force (AATF) [3]. The report contains a variety of data elements (for demographics, laboratory tests, etc.) We focus on the data elements presented in table 7 of this report under the subheading “Outcomes” [3].

The evaluation proceeds as follows: For each AATF recommended outcome we use Jaccard-based string similarity measure to identify the 5 closest matches from the CTG outcomes. The AATF outcomes, along with the matching CTG outcomes are written out to a spreadsheet for the purpose of manual review. The manual reviewer then selects the single best match from the 5 matches automatically identified, and assigns one of three quality grades to the match: “Exact”, “Near match”, “No match”. The “Near match” label is used when the best counterpart we could identify from CTG is slightly less or more specific than the AATF outcome, or has the same outcome domain but using a different measure.

### 3. Results

#### 3.1. Implementation and Use of the Pipeline

We implemented the pipeline using Python 3.7. We used the built-in *urllib* library for the submission of the API requests and we used *BeautifulSoup* library for the parsing of the XML. We ran the pipeline on a list of two conditions: *Acute Coronary Syndrome* and *Coronary Artery Disease*. ClinicalTrials.gov's API identified 1055 trials related to the former and 7401 trials related to the latter. 1,016 were common to both result sets, giving us a total of 7,440 unique trials after aggregating the results for both conditions. The pipeline fetched the XML data for each trial. 576 trials (8%) listed no outcomes. From the remaining 6,864 trials, the pipeline parsed 8,188 primary outcomes and 19,341 secondary outcomes.

#### 3.2. Evaluating the Coverage of the Pipeline

For outcome comparison, we grouped the 31 outcome elements recommended in the AATF report into 23 domains by grouping dates of events with the events (e.g. *Bleeding event*, *bleeding location*, *date of bleeding* are grouped together, so are *Death* and *date and time of death*). Then, with each outcome string represented as a bag of words, the pipeline uses Jaccard similarity to identify the 5 best matches from CTG data for each ACCF/AHA outcome. The best match is then manually selected from 5 automatically identified. We found that 17 of 23 (74%) outcomes had exact matches in the CTG identified outcomes, while the remaining 6 outcome domains (26%) had near matches. 100% of the outcome domains had matches in CTG data. Table 1 shows the mapping of expert recommended outcomes to the best counterparts identified from CTG.

We then proceed to consider the forward usefulness of the pipeline. That is, how many of the expert recommended outcomes does the pipeline identify when restricted to analyzing only trials posted prior to the publication of the report (January 2013)? We find that 21 out of the 23 outcomes had matches in CTG data from trials that were posted to CTG prior to January 2013. For the 2 outcomes where the earliest came from trials following the publication of the report: *Cardiac rupture/ventricular septal defect*, and *Supraventricular tachycardia*, we are able to find a near-match for the latter from pre-2013 trials (Outcome: "Number of patients with and without depression suffering from ventricular and/or supraventricular arrhythmias", from trial: NCT00622024). Putting the coverage of the pipeline when restricted to pre-report trials at 22/23 (96%).

### 4. Discussion

The automated development of core outcome sets is an important task for research informatics that has not yet fully benefitted from the recent trends in big data and natural language processing. The tremendous amount of detailed data available from ClinicalTrials.gov have found a variety of innovative uses for other research tasks. For example, Huser and Cimino have worked to link CTG to PubMed to analyze the proportion of trials that reported results through publication [4] and to understand the quality and completeness of the links [5]. Anderson et al. used CTG data to study level of compliance with result reporting requirements [6]. When it comes to outcomes and

**Table 1.** Evaluating the coverage of outcomes identified by our pipeline against the outcomes recommended by ACCF/AHA for acute coronary syndromes and coronary artery disease.

Recommended Outcome by AATF	Best Matching Outcome from CTG	Match Quality
Death / Date and time of death	Death	Exact
Acute Myocardial Infarction	Acute Myocardial Infarction	Exact
Recurrent Myocardial Infarction	Recurrent Myocardial Infarction	Exact
Reinfarction date	Reinfarction	Exact
Recurrent rest angina with/without electrocardiographic changes	Recurrent angina	Exact
Unstable angina requiring hospitalization	Unstable angina requiring hospitalization	Exact
Heart failure/date	Heart failure	Near match
Cardiogenic Shock/Date	Cardiogenic shock	Exact
Stroke/date/type	Stroke, Stroke Type	Exact
Bleeding (TIMI major, TIMI minor, or none)	Bleeding TIMI major, minor and combination	Exact
GUSTO bleeding classification	GUSTO severe or moderate bleeding	Near match
Bleeding event/ Location of bleeding/date	Bleeding event	Exact
Surgical or procedural intervention	Procedural complications	Exact
Transfusion	Blood transfusion	Near match
Units of blood given	Transfusion Units	Near match
Date of first RBC transfusion	RBC Transfusion (volume and rate)	Near match
RBC transfusion related to CABG	RBC Transfusion (volume and rate)	Exact
Thrombocytopenia	Thrombocytopenia	Exact
Cardiac rupture/ventricular septal defect	Ventricular septal defect(VSD) recurrence	Near match
Atrial arrhythmia	Atrial arrhythmia burden	Exact
Supraventricular tachycardia	Arrhythmias (ventricular and supraventricular)	Near match
Ventricular arrhythmia	ventricular arrhythmia	Exact
High-degree AV block	Incidence of second- or third-degree AV block	Near match

common data elements, Luo et al. proposed a semi-automatic approach for identifying inclusion criteria CDEs [7]. Our findings are congruent with the previous study focusing on core outcome sets in chronic lung disease [8]. Compared to a set of gold-standard outcomes recommended by a taskforce of experts, our pipeline identifies exact or near-matches for 100% of the recommended outcome set. When the data is restricted to trials posted prior to the AATF report’s publication, the coverage of the pipeline results degrade only slightly, to 96%. We believe this demonstrates the great potential of this approach in aiding and informing the development of clinical outcomes.

The next steps will address the large size of the results generated by CTG queries. Manual inspection indicates that a significant portion of the variety of the outcome sets is due to variation in the word choice or phrasing. Mapping to standard terminologies [9] such as UMLS might be useful in this regard and it will be the focus of our future work.

## References

- [1] Thurmond VA, Hicks R, Gleason T et al. Advancing integrated research in psychological health and traumatic brain injury: Common data elements. *Archives of physical medicine and rehabilitation*. 2010;91(11):1633-1636.
- [2] Sheehan J, Hirschfeld S, Foster E et al. Improving the value of clinical research through the use of Common Data Elements. *Clinical Trials*. 2016;13(6):671-676.
- [3] Cannon CP, Brindis RG, Chaitman BR et al. 2013 ACCF/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes and coronary artery disease: A report of the American College of Cardiology Foundation/American Heart Association Task Force on clinical data standards (writing committee to develop acute coronary syndromes and coronary artery disease clinical data standards). *Journal of the American College of Cardiology*. 2013;61(9):992-1025.
- [4] Huser V, Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PloS one*. 2013;8(7):e68409.
- [5] Huser V, Cimino JJ. Precision and negative predictive value of links between ClinicalTrials.gov and PubMed. *AMIA Annu Symp Proc*. 2012;2012:400-8.
- [6] Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. *New England Journal of Medicine*. 2012;372(11):1031-1039.
- [7] Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *Journal of biomedical informatics*. 2013;46(1):33-39.
- [8] Elghafari A, Finkelstein J. automated identification of common disease-specific outcomes for comparative effectiveness research using ClinicalTrials.gov: Algorithm development and validation study. *JMIR Med Inform*. 2021;9(2):e18298.
- [9] Elghafari A, Finkelstein J. Introducing an ontology-driven pipeline for the identification of common data elements. *Stud Health Technol Inform*. 2020 Jun 26;272:379-382.