Set of derived metadata elements for comparing cohorts from human clinical trials datasets and EHR

Vojtech Huser, Craig Mayer, Nick Williams, Sigfried Gold, Kin Wah Fung

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

We **converted** several HIV **trials** and one HIV EHR cohort **into OMOP** and designed set of measures (called Scyros) that allow data density comparison.

Introduction

Clinical trials are increasingly being conducted close to routine healthcare taking advantage of consented data import from Electronic Heath Records (EHR) in order to reduce costs. There is also a trend towards converging data formats used by clinical trials (e.g., Clinical Data Interchange Consortium (CDISC) formats or REDCap data dictionary (DD) format) and data formats used by observational databases with EHR data (e.g., Observational Medical Outcomes Partnership (OMOP) common data model or Fast Healthcare Interoperability Resources (FHIR) formats).[1]

In light of this harmonization, we study how limited data conversion of clinical trial data into simplified OMOP model format can facilitate comparison between different trial data cohorts as well as the comparison between trial data cohorts and corresponding observational EHR patient cohorts.

Methods

First, we obtained data from three human clinical trials (in HIV domain) from several data sharing platforms and trial networks. We use HIV trials because of a larger project focused on Common Data Elements (CDE) in HIV; however, the methodology developed below translates to other medical domains. We defined psOMOP model (ps stands for pragmatic and study; pragmatic because it is only a small subset of OMOP and study because it is meant for research studies). psOMOP is based on full OMOP Common Data Model (CDM) version 6. Second, we measured the time spent converting trial data into psOMOP from a non-standard format. Unlike our previous work [1], we specifically targeted trials that did not use any standard format to share de-identified individual participant data (IPD). We recoded separately time spent on DD review, data loading and psOMOP data transformation. Third, we created a set of data characterizations measures (we refer to this set as Scyros snapshot). Scyros was designed to use the same approach to data characterizations as the Achilles tool developed by the Observational Health Data Science and Informatics (OHDSI) community. In some Scyros measures, there was direct correspondence to Achilles; however, Scyros calculations use R dplyr package data query logic rather than SQL. Finally, to demonstrate the comparison of trial and EHR data, we computed the same measures on EHR cohort of HIV patients (converted from PCORNet model into psOMOP model; through collaboration with Great Plains Collaborative). As an additional research cohort, we generated another HIV patient cohort from the NIH Clinical Center research data warehouse (called BTRIS).

Results

We obtained data from the following trials: (1) HPTN068 trial (ClinicalTrials.gov id: NCT01233531); (2) ATN109 (NCT01751646), (3) PHACS (NCT01418014). The mean data transformation times were 2.31 hours for DD analysis, 1.03 hours for data loading and 0.59 hours for data conversion.

We only considered the PERSON, VISIT_OCURRENCE and MEASUREMENT tables. The psPERSON table consisted of columns person_id and year_of_birth. The psVISIT_OCCURRENCE table consisted of columns person_id and visit_start_date. We designed a total of 24 measures grouped into 8 higher level constructs. For example, mean number of visits per participant allows comparison of data density (and visit temporal pattern) across studies or patient cohorts. The full list is available in file S1-measures-description at our project repository at https://github.com/lhncbc/CDE/tree/master/scyros including links (for 7 of them where link exist) to a corresponding Achilles measure. File S2-comparison shows the final comparison of the four research cohorts and one EHR cohort.

Discussion

For data re-using researchers, Scyros snapshot can be used to identify datasets most suitable for their research (assuming at least some data characteristics important for picking the dataset are covered by Scyros). We have piloted post-hoc conversion of clinical study data into a highly pragmatic and simplified subset of an EHR CDM. The majority of time needed for this conversion was spent in DD and data loading stage. Since many sharing platforms have dedicated staff and processes for post-hoc data processing that inherently require DD analysis and data loading, generation of psPERSON and psVISIT_OCCURRENCE tables can possibly be added to this step.

With respect to input data, for one trial in our sample, visit dates were relative and provided as days-since-birth. Because relative dates are not natively supported by OMOP, we imperfectly calculated absolute visit dates from year of birth (hypothetically assigning July 2nd as date of birth for every participant). Our OMOP transformation process was thus able to compensate for this lack of support in the model. Another trial-data specific factor we noted was related to the nature of Scyros measures. Some Scyros measures are absolute dates, such as ‘earliest visit date’ and ‘latest visit date’. These provide data re-using researchers an important indication about data currency for their intended use. We observed, that it is possible to de-identify the trial dataset in such a fashion, that absolute temporal measures cannot be determined from the de-identified dataset. (use of age at enrollment; redaction of year of birth and strict use of relative visit dates). Scyros computation centrally by the sharing platform or data depositor (prior to masking the true dates) can avoid this problem. Our work has several limitations. First, we used only a small set of trials that come from a single disease domain of HIV. Second, our measurement of time spent on data conversion rely on a single analyst performing the conversion.

References

1. Huser V. Converting clinical trial data between CDISC SDTM and OMOP CDM OHDSI Symposium 2018

2. Smith CT, Hopkins C, Sydes M, et al. Good practice principles for sharing individual participant data from publicly funded clinical trials. Trials 2015; 16: O1.

Exhibit 1: Selected Scyros measures results for five compared datasets

trial ID	-	NCT01233531	-	NCT01751646	NCT01418014
acronym	GPC	HPTN068	BTRIS	ATN109	PHACS
n	2032	2533	4040	102	678
span	3206	1563	15302	844	3206
enroll_age_median	45	15	58	23.03287671	11
enroll_age_max	83	21	67	25.3260274	2.620903313
enroll_age_min	10	12	25	18.53424658	16
end_age_median	48	18	61	23.63287671	16
end_age_mean	48	18	61	23.57206022	16.30684932
end_age_max	85	24	67	26.0739726	5.893887156
end_age_min	10	13	35	19.22465753	25
span_median	811	910	0.003877315	331	731
span_min	0	0	0	40.0000032	0
span_max	3159	1488	0.154583333	509.0000256	822
visit_count_median	11	4	3	3	5
visit_count_mean	24	4	5	3	5

Acknowledgements: This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by the NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services. **Contact:** vojtech.huser@nih.gov

NIH U.S. National Library of Medicine