Published in final edited form as:
Methods Inf Med. 2022 May ; 61(1-02): 3–10. doi:10.1055/s-0041-1739361.

A Semi-Automated Term Harmonization Pipeline Applied to Pulmonary Arterial Hypertension Clinical Trials

Ryan J. Urbanowicz[1], **John H. Holmes**[1], **Dina Appleby**[1], **Vanamala Narasimhan**[2], **Stephen Durborow**[1], **Nadine Al-Naamani**[3], **Melissa Fernando**[1], **Steven M. Kawut**[3]

[1]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States

[3]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States

Abstract

Objective—Data harmonization is essential to integrate individual participant data from multiple sites, time periods, and trials for meta-analysis. The process of mapping terms and phrases to an ontology is complicated by typographic errors, abbreviations, truncation, and plurality. We sought to harmonize medical history (MH) and adverse events (AE) term records across 21 randomized clinical trials in pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension.

Methods—We developed and applied a semi-automated harmonization pipeline for use with domain-expert annotators to resolve ambiguous term mappings using exact and fuzzy matching. We summarized MH and AE term mapping success, including map quality measures, and imputation of a generalizing term hierarchy as defined by the applied Medical Dictionary for Regulatory Activities (MedDRA) ontology standard.

Results—Over 99.6% of both MH (N = 37,105) and AE (N = 58,170) records were successfully mapped to MedDRA low-level terms. Automated exact matching accounted for 74.9% of MH and 85.5% of AE mappings. Term recommendations from fuzzy matching in the pipeline facilitated annotator mapping of the remaining 24.9% of MH and 13.8% of AE records. Imputation of the generalized MedDRA term hierarchy was unambiguous in 85.2% of high-level terms, 99.4% of high-level group terms, and 99.5% of system organ class in MH, and 75% of high-level terms, 98.3% of high-level group terms, and 98.4% of system organ class in AE.

Conclusion—This pipeline dramatically reduced the burden of manual annotation for MH and AE term harmonization and could be adapted to other data integration efforts.

Address for correspondence Ryan J. Urbanowicz, PhD, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 403 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, United States (ryanurb@pennmedicine.upenn.edu).

Conflict of Interest
None declared.

Keywords

adverse events; medical history; pulmonary arterial hypertension; informatics

Introduction

The power of statistical and machine-learning analyses is ultimately limited by sample size.[1] Biomedical studies and clinical trials enroll a sample size based on the power of detecting clinically important differences between groups.[2] Particularly in rare diseases, individual participant data meta-analysis can increase the power and generalizability of both the main and secondary aims.[3] The analysis of data can be standardized and performed appropriately across all studies. Outcomes of interest which are not presented in the publications of the individual studies can be examined. Prognostic models can be derived and/or validated in the combined study samples. The effect of a drug (or class of drugs) can be studied in a subgroup of patients, to determine if there is a “responder” population or, conversely, a population which suffers a particularly high rate of side effects.[4]

Individual participant data meta-analyses also pose distinct challenges outside of standard analytical challenges such as considering potential confounders. Harmonizing many thousands of data fields between multiple studies is extremely work intensive.[5] Distinct data sources can vary at the global level (e.g., included variables, how named and organized) or local variable level (e.g., feature values encoded differently, different units of measurement, and text that does not always correspond to a standardized semantic mapping). Several strategies, tools, and pipelines have been proposed to address data harmonization at different levels, phases of data collection and processing, and that target different data domains. Examples include global harmonization of detailed clinical models for clinical study data standards,[6] a pipeline to facilitate the collection of standardized medical metadata rather than deal with posthoc harmonization,[7] a flexible platform to facilitate the basic, high-level harmonization of individual patient data for meta-analysis relying on well-defined domain-specific data dictionaries,[8] and a variety of other efforts designed primarily to facilitate generalized data harmonization at the global level.[9,10] Data harmonization efforts like these seek to produce an integrated file with unified semantics for all features and feature values.

Beyond these global efforts, “term” harmonization of text (e.g., words, and phrases) at the local variable level is challenged by typographical errors, spelling errors, abbreviations, multiple languages, slang, truncation, plurality, phrasing differences, and terms that are missing from the target terminology standard. Numerous efforts to overcome these problems have largely focused on ontology development with the goal of mapping terms to a standard vocabulary that enforces syntactic and semantic harmonization between two or more heterogeneous sources of data. Several prominent initiatives have addressed harmonization through the establishment of common data models including the Patient-Centered Clinical Research Network,[11] the Observational Medical Outcomes Partnership, the common data model used in the Observational Health Data Sciences and Informatics network,[12] the Informatics for Integrating Biology and the Bedside,[13] and the Fast Healthcare Interoperability Resources specification[14] which includes a modular approach to data standardization. The Clinical Data Interchange Standards Consortium (CDISC)[15] also supports a variety of tools for standardizing the planning, collection, organization, and analysis of clinical trial data.[16–19] However, none of these alone specifically address the logistical demands of harmonizing text-based variables in complex clinical trial data. Time-consuming manual effort from trained experts is typically required to map heterogeneous terms in the target data to a terminology standard. However, this process can be facilitated with the application of text-matching strategies, including approximate, i.e., "fuzzy," string matching methodologies[20] as well advanced machine-learning approaches that require prior training for specific use cases.[21]

Objectives

In this project, we established a procedure to standardize and automate the process of term harmonization across multiple data sources applied here to 21 randomized clinical trials (RCTs) of medical treatment for pulmonary arterial hypertension (PAH) or chronic thromboembolic pulmonary hypertension (CTEPH), and a subset of open-label (OL) extensions. We focused on term harmonization for adverse events (AEs) and medical history (MH) data since the nature of these variables provided the best opportunity and need for automated harmonization within the larger task of harmonizing all data types across the RCTs. Specifically, mapping these terms (or sequence of multiple terms) was challenging due to (1) inconsistency of documentation across centers, trials, and calendar time, (2) errors from manual data entry transcription from natural language text, (3) the inconsistent use of a specific terminology/ontology standard across studies, (4) a large variety of values existing for both AE and MH, and (5) the selected term/ontology standard, i.e., Medical Dictionary for Regulatory Activities (MedDRA)[22] including over 70,000 possible specific terms at the lowest level of the term hierarchy.

We present the AE and MH term harmonization procedure distinguishing all aspects automated by our pipeline and show the results after applying this harmonization procedure to the 21 RCTs of PAH, highlighting lowest level term (LLT) mapping success rates, quality scores, and imputation evaluation statistics for preferred term (PT), high level term (HLT), high level group term (HLGT), and system organ class (SOC).

Methods

Here we (1) summarize the target RCT data and ontological standard, (2) summarize the data model and global plan for variable harmonization, (3) detail the pipeline steps for mapping the LLT for AE or MH, and (4) detail the pipeline steps for imputing and mapping the respective term hierarchy (i.e., PT, HLT, HLGT, and SOC).

Target Data and Ontological Standard

This project focused on the harmonization of 21 (mostly phase 3) RCTs of patients with PAH and CTEPH completed at different sites throughout the world between 2000 and 2015. A total of 11 candidate treatments were examined across the 21 trials. Additionally, seven of these trials include their respective OL extensions, yielding a total of 28 studies for harmonization. These studies are detailed in supplementary Table S1 (available in the online version only) along with a study ID, a unique trial name, phase (i.e., 2, 3, or OL), treatment, and associated publications.[23–46] The data from these trials were provided to us by the Cardiorenal Division of the Food and Drug Administration and are not publicly available. From the perspective of term harmonization, these studies were heterogeneous with respect to (1) what variables were collected, (2) how they were organized, (3) units of measurement, (4) use of slang or abbreviations, (5) which terminology standards or versions (if any) were applied in documenting the values for certain variables, (6) whether raw or analyzable datasets were available, and (7) the overall data model used to organize the respective database. More generally, these studies were also heterogeneous in many other ways including the PAH-specific background therapy (i.e., treatment naive vs. double or even triple therapy), time frame of studies, and definition of AE per study.

We focused this analysis on the term harmonization of AEs and MH records often with a “one-to-many” patient-record mapping. In some (but not all) of the 28 studies, an unspecified version of MedDRA was applied in documenting AE and MH values for LLT, and in some studies and subjects, PT, HLT, HLGT, and SOC were also documented. MedDRA is a rich and highly specific standardized medical terminology and ontology that includes terms at varying degrees of generality related to diseases, diagnoses, signs, symptoms, medical/surgical procedures, family history, and therapeutic indications.[22] MedDRA is also widely used and regularly updated. In the present project, we utilized terms defined as “current” within MedDRA v21.0 as our mapping standard.

Global Data Model and Harmonization

The overall harmonization of these trials adopted the standard data tabulation model, a CDISC standard,[47] that is well documented and accepted by the pharmaceutical industry, pharmacoepidemiology, and other clinical research areas. It provides excellent foundational domains and variable names; however, it provides no standards for coding quantitative or categorical variable values, or terminology standards as needed to harmonize AE and MH.

Term Harmonization Pipeline Overview

This semi-automated term harmonization pipeline included (1) data preprocessing, (2) exact matching, (3) fuzzy matching, and (4) integration and quality control (QC) of expert manual annotations. This was first completed for the most specific LLT as defined by the MedDRA standard,[22] and then using a combination of imputation and term mapping for each of the more general term levels in the MedDRA hierarchy, i.e., PT, HLT, HLGT, and SOC. We also applied a mapping quality metric to track the quality and fidelity of term harmonization when there was any ambiguity.

Lowest Level Term Harmonization

The first part of the term harmonization pipeline focused on harmonizing the most specific AE or MH term/phrase available, referred to by MedDRA as the LLT. For simplicity, here we will primarily focus on describing MH mapping; however, this procedure is identical for both MH and AE except for how variables are named, and the availability of certain accessory variables in mapping the respective variable categories.

Preprocessing—The first stage of LLT harmonization was to generate a combined file containing the MH data from all 28 studies where each record represents a target MH term (referred to here as the "primary term" [PRT]) attributed to a given subject from each study. Additional subject and study information was preserved in this file for future reference including any other higher level MH term information associated with the PRT that may have been collected whether it was in line with the MedDRA term hierarchy or not. For a limited number of records in the combined clinical file, LLT, PT, HLT, HLGT, and SOC term information was also available. However, different versions of MedDRA term standards had been used which may have led to a marginal number of inconsistencies, and often errors were found in records making these preliminary mappings unreliable.

Next, we conducted exploratory analyses of the combined file, assessing the number of total and unique terms to be mapped, as well as variable missingness. Any records missing a study-derived PRT were dropped from further consideration (i.e., unmapped). We began with 37,105 MH records, reduced to 37,083 after dropping records with no PRT, which included a total of 21,452 unique MH entries to be mapped. For AE, we began with 59,084 records, reduced to 58,170 after dropping records with no PRT, which included a total of 20,824 unique AE entries to be mapped. To reduce the mapping effort required, we temporarily dropped all but one of each unique PRT record but included all records in the final mapped file as reflected in the results below.

Next, we conducted exploratory analysis of the respective standard terminology files (i.e., MedDRA term files for LLT, PT, HLT, HLGT, and SOC). LLT included a total of 78,808 terms. Filtering out terms that were not "'current" (as defined by MedDRA[22]) yielded 69,531 unique LLTs to which our MH or AE terms would be mapped. Similarly, PT included 23,088 unique terms, HLT included 1,737 unique terms, HLGT included 337 unique terms, and SOC included 27 unique terms.

Exact Matching—The second stage of LLT harmonization involved identifying records with terms that exactly matched any term/phrase found in MedDRA's LLTs. These yielded mapped terms with the highest confidence, and further reduced the amount of computing and manual annotation time required for downstream. Python's "casefold" method was applied to conduct "caseless" exact text matching. Exact matching was initially applied to terms in the PRT column, and then subsequently to any available LLT information in the combined file to increase the chances of identifying an exact match. If an exact match was still not found, we similarly checked for exact MedDRA LLT matches using any additional term information that may have been available for that record. If an exact match was found, the corresponding standardized MedDRA term was added to a new "mapped term" column, and the corresponding MedDRA term code similarly added to a "mapped term code" column. These codes were critical to imputing the term hierarchy later in this procedure.

During this and later stages of mapping, we also added a "map quality score" column to the file. For MH, a match received a quality score of 0 if the exact match was found using both the PRT and LLT columns (consensus found). Score 1 was assigned if the exact match occurred in the PRT column alone, score 2 was assigned if it was in the LLT column alone, and score 3 was assigned if it occurred in the accessory term column. Similar quality scores could be customized to the needs of a given term harmonization task, and they allowed for a more detailed level of map-quality evaluation as well as improved confidence and reproducibility in the mapping procedure.

Fuzzy Matching—The third stage of LLT harmonization addressed mapping of any terms unresolved by the exact matching stage. Given the unique constraints of our MH and AE harmonization problem, there was no reliable strategy to completely automate the mapping of terms that did not exactly match. This was because even a single-letter difference had the potential to completely change the underlying meaning of a given text value. It was also due to the many text mapping challenges outlined above, including there being a large number of possible terms to be mapped to. Other harmonization tasks may be easier to completely automate.

Our pipeline adopted a specific form of fuzzy matching. Fuzzy matching generally involves estimating the degree of match between individual words or sentence level segments of text. Fuzzy matching can rely on a variety of distance metrics. For example, Levenshtein distance estimates distance between two words based on the minimum number of single-character edits required to change one word into the other. Differently, phonetic algorithms such as SoundX[48] can be effective at detecting homophones, but can oversimplify matching in complex, large-scale text matching. Fuzzy matching distances can also be applied in ways that directly compares (1) all text, i.e., *simple*, (2) the best matching length-m substring, i.e., *partial*, (3) all text, after sorting all words alphabetically, i.e., *token sort*, and (4) all text, after first sorting words found in both alphabetically and then adding any other words alphabetically, i.e., *token set*. Preliminary testing of performance and computational time led us to adopt the “fuzzywuzzy” Python package which uses the Levenshtein distance,[49] and the “simple” approach. Specifically, in all fuzzy matching evaluations, the fuzzywuzzy “extract” method was applied which outputs the top five simple MedDRA LLTs for each unresolved term instance. These potential matches and their scores were added to the mapping file.

Manual Mapping by Domain Expert(s)—The fourth stage of LLT harmonization required manual inspection and mapping. The preceding fuzzy matching stage was not always reliable (particularly in large-scale complex mapping tasks), making manual mapping a necessity to maintain high mapping accuracy. However, fuzzy matching facilitated this process by presenting medically trained mappers with the top five matching MedDRA terms as recommendations. If none of these options were accurate, the mapper resorted to applying their own knowledge and searched for an appropriate term in MedDRA manually.

During this process, the mappers could either select the best of the five terms selected by the fuzzy match process (identified by their index number 1 to 5), or copy an appropriate MedDRA term into the “mapped term” column. They also assigned the “map quality score,” e.g., for MH, a quality score of 4 was assigned for “high-confidence” fuzzy mappings, and 5 for “medium-confidence” fuzzy mappings. A score of 6 was assigned to indicate that the term had been examined but no clear match had been found.

Two clinical research coordinators, nurses, or physician fellows conducted a first pass of mapping on respective, random, and approximately equal record subsets. An attending physician conducted a second pass of all records, verifying or correcting terms given a score of 5 and attempting to add mappings of terms given a score of 6. This process of manual mapping benefited from multiple passes to improve the overall coverage and quality of term harmonization. We provided standardized documentation and training for mappers before their review (refer Supplementary Materials, available in the online version only).

Annotation Merging and Quality Control Checks—The fifth stage of LLT harmonization began by merging the manually mapped files (if they had been split into subsets to accommodate multiple mappers in the preceding stage). Next, our pipeline automated QC procedures. This included checking that: (1) the "mapped term" either included copied text that exactly matched a term in MedDRA LLTs or a mapper-selected fuzzy term index (1–5), (2) only records with a quality code of 6 were missing a mapped term entry, and (3) all term quality scores were valid (i.e., 0–6). During this process, fuzzy term indexes (1–5) previously placed in the "mapped term" column were replaced with the respective MedDRA LLT term, and the corresponding MedDRA term code was added. Following a round of QC, flagged records with issues were addressed by an expert, and a subsequent QC round applied to confirm success. If needed, another cycle of fuzzy matching (with a different algorithm), manual annotation, and QC checks could be applied to minimize the number of terms left unresolved.

Term Hierarchy Mapping

The last phase of the term harmonization pipeline shifted from LLTs to mapping the more general terms of the MedDRA hierarchy (i.e., PT, HLT, HLGT, and SOC) allowing the consolidation of possible states/values for analyses. Consider that within the original MH LLTs from the drug trials, 21,452 of the 37,083 were unique. With each term occurring less than twice on average, it would be difficult to leverage these more specific terms to characterize the populations in the trials. It may be easier to phenotype patient subsets using more general term categories than more specific ones.

With the LLT-mapped terms and codes in place, this phase sought to either impute (if no MedDRA hierarchy information was available) or apply exact/fuzzy matching (if it was available) to map the four increasingly general MedDRA term categories. MedDRA includes ontology files defining the term connections between each level of the hierarchy (i.e., LLT/PT, PT/HLT, HLT/HLGT, and HLGT/SOC). These are captured using the aforementioned term codes. The first level of imputation, i.e., LLT/PT, is a simple direct mapping, i.e., each LLT is connected to a single PT. This mapping was completely automated and deterministic. The corresponding PT term and term code were added to the map file. For the remaining levels, there could be one or more possible term connections to choose from (i.e., branches, or "secondary SOC"). Whenever only a single branch was present, that corresponding term was automatically chosen. When multiple branches were available, the pipeline attempted to use any available term hierarchy information from the original study data to pick the matching or closest matching option (discussed below). If no term hierarchy information was available from the original study, our pipeline selected the first branch found.

Assuming term hierarchy information was available, mapping of the best HLT, HLGT, or SOC branch proceeded similarly to the process described for LLTs. Specifically, exact matching was first attempted, followed by fuzzy matching identification of the five best matches if exact matching failed. As before, fuzzy matching was followed by manual selection of the best option by a mapper. If no best match could be identified manually, the first branch was selected. Notably, exact and fuzzy matching was no longer conducted against all MedDRA terms, but only using terms identified as candidate branches, making it much faster. The overall process of imputing and mapping PT, HLT, HLGT, and SOC across all instances must be completed sequentially.

Similar to LLT mapping, each term hierarchy mapping included automatic or manual expert annotation of map quality scores. Here, the highest quality score of 1 indicated a deterministic mapping to a single branch. Score 2 indicated resolution with an exact match, score 3 indicated fuzzy match resolution, score 4 indicated default selection of the first available branch, and score 5 indicated that no term could be chosen (when the preceding more specific term was unavailable).

Next, this phase underwent a similar merging and QC check before yielding a final map file, linking the original data records to standard LLT MedDRA terms, along with the corresponding term hierarchy and map quality scores for each. Lastly, the pipeline outputted mapping summary statistics for all term levels as reported in the Results section. The entirety of the MH semi-automated harmonization pipeline is implemented as a set of Python-based Jupyter notebooks available at: https://github.com/UrbsLab/auto_term_harm_pipe.

Ethical Considerations

Studies using this dataset have been determined to be exempt from review by the Institutional Review Board of the University of Pennsylvania.

Results

In this section we present the results of applying this harmonization pipeline to both MH and AE terms. Table 1 summarizes the total, exact, fuzzy, and unmapped counts across the 28 studies. Less than 0.4% remained unmapped for both MH and AE terms following application of this term harmonization pipeline (including records dropped in the first pass). Unmapped records included both those removed for having no PRT and those where no clear MedDRA term mapping could be determined. Closer inspection of these unmapped records confirmed that they were nearly all partial words or unintelligible phrases. Tables 2 and 3 break down these term mappings further based on the automatically or manually assigned quality codes. Keeping in mind that exact matches were completely automated, but fuzzy matches required manual annotation, the “high,” “medium,” or “unmapped” confidence scores for fuzzy matching reflect the success of the manual efforts that were facilitated by automated fuzzy matching. For both MH and AE, the vast majority of terms were exactly matched or manually mapped after fuzzy matching with a high confidence.

Lastly, Table 4 breaks down the counts and quality scores for the term hierarchy mapping. PT is not included in this table because all PT terms were mapped deterministically from the mapped LLTs (i.e., they all have an imputation quality score of 1). For MH, 85.2% of HLTs, 99.4% of HLGTs, and 99.5% of SOCs were mapped without ambiguity, i.e., exact or fuzzy matching was able to be completed based on available information in the original studies. For AE, these statistics were 75% for HLTs, 98.3% for HLGTs, and 98.4% for SOCs. These values allowed us to gauge reliability of these mappings.

Notably, the automated elements of this pipeline (run in stages) took roughly 3 to 4 days of compute time on a single PC workstation with 3.49 GHz processor and 64 GB of RAM for both AE and MH, respectively.

Discussion

This work sought to illustrate a rigorous procedure and offer publicly available code that can be adapted to automate the process of term harmonization in other domains where an ontological standard is available. This pipeline successfully automated exact matching of 75 to 85% of MH or AE terms respectively, which eliminated the need for manual checks of these terms. There is no reason this step should be completed manually, as it is an obvious time waste. Fuzzy matching and manual annotation on the remaining terms can be relatively expensive in both computing and expert mapping time, where saving the latter is our primary concern. While it was impractical to reproduce the tedious process of harmonization without fuzzy matching to compare mapping times, we can extrapolate the estimated time savings by examining high versus medium confidence fuzzy matching in Tables 2 and 3 (MH and AE, respectively). High confidence matches occurred when the manual annotator was able to quickly and easily select a “best” matching term from the five top matches identified by fuzzy matching. Our annotators estimated this took an average of 4 seconds per term. Medium confidence matches indicate that the annotator found none of the proposed fuzzy matches to be appropriate. At this point the annotator would apply their own expertise to search MedDRA manually for an appropriate match. Our annotators conservatively estimated this to take an average of 30 seconds per term. Therefore, we estimate that MH harmonization of remaining (i.e., not exactly matched) LLTs took 10.5 hours of manual annotation time with automated fuzzy matching, but would have taken 77 hours without. For AE terms we similarly estimate it took 16.9 hours with fuzzy matching and would have taken 68.8 hours without. Furthermore, manually copying and pasting terms between the MedDRA standard and the harmonized file leaves greater opportunity for errors than annotators being able to select an appropriate option from a list of five on the same worksheet. Overall, application of this pipeline greatly reduced the manual workload required for this MH and AE harmonization task, as well as ensured fidelity of our final mapping through the use of automated QC checks.

The semi-automated pipeline developed for this study provides an adaptable example of how to conduct term harmonization with exact matching, fuzzy matching (with subsequent manual annotation), and generation of a term hierarchy (if such a hierarchy is available within the target ontology). However, this approach still has some limitations. First, while we believe we have automated this procedure as much as is possible, a significant amount of manual effort may still be required to apply this pipeline to other tasks, particularly if the frequency of exact matches in a given harmonization task is significantly lower than in the present study. Second, this pipeline is only practical for application to single words or a relatively small sequence of words, as defined by a target ontology. It is not applicable to full sentences or other larger groups of text. Third, when information was not available in the records to facilitate selection of the most appropriate branch in the hierarchy mapping (section “Lowest Level Term Harmonization”), the first option was selected by default, which likely leads to some bias in the assignment of these more general terms. Lastly, with respect to the MedDRA ontology, not all laboratory tests can be adequately represented by the binary MedDRA categorization (i.e., normal vs. abnormal) or other LLT or PT codes. While not an issue in this study, this could be a considerable limitation in other areas.

Future work will (1) extend this pipeline to harmonize concomitant medications using RxNorm as our terminology standard, (2) explore the reliability of our manual annotation procedure by comparing mappings with a different group of mappers, and ultimately (3) leverage the larger sample size of the fully harmonized database to conduct statistical and machine-learning analyses of target subject outcomes.

Clinical or Public Health Implications

This new method for harmonizing MH and AE data from multiple RCTs performed over several years could facilitate individual participant data meta-analyses in other areas.

Conclusions

In this project, we have proposed an efficient procedure for conducting term harmonization for integrating MH and AE variables across 21 RTCs including a total of 28 studies.
In the process we developed a freely available semi-automated harmonization pipeline implemented over a set of Python-based Jupyter notebooks. This procedure included map quality scores to trace the ultimate mapping success and allow downstream investigators that use this harmonized database to go back and estimate data reliability. We reported summary statistics breaking down the frequency with which we were able to rely on exact matching, fuzzy matching combined with manual mapping, and unmappable terms. We found evidence of high fidelity in our resulting MH and AE term harmonization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Mamta Patel, Kathleen Boyle, Randi Goodman, Tess LaPatra, Sai Narasimmal, Katelyn Pierce, Jasleen Minhas, and Jeff Min for their efforts in conducting the first pass of the MH and AE manual mapping procedures.

Funding

This work was supported by Cardiovascular Medical Research and Education Fund (CMREF), Aldrighetti Research Award for Young Investigators, and NIH 2K24HL103844–06 and 5K23HL141584–03.

References

1. Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol 2008;59(01):537–563 [PubMed: 17937603]
2. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2(02):93–113 [PubMed: 7273794]
3. Evans JDW, Girerd B, Montani D, et al. BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. Lancet Respir Med 2016;4(02):129–137 [PubMed: 26795434]
4. Halliday SJ, Hemnes AR. Identifying “super responders” in pulmonary arterial hypertension. Pulm Circ 2017;7(02):300–311 [PubMed: 28597766]
5. Lee JS-H, Kibbe WA, Grossman RL. Data harmonization for a molecularly driven health system. Cell 2018;174(05):1045–1048 [PubMed: 30142341]
6. Jiang G, Evans J, Oniki T, et al. Harmonization of detailed clinical models with clinical study data standard. Methods Inf Med 2015;54(01):65–74 [PubMed: 25426730]
7. Kock-Schoppenhauer A-K, Kroll B, Lambarki M, et al. One step away from technology but one step towards domain experts-MDRBridge: a template-based ISO 11179-compliant metadata processing pipeline. Methods Inf Med 2019;58(S 02):e72–e79 [PubMed: 31853911]
8. Kalter J, Sweegers MG, Verdonck-de Leeuw IM, Brug J, Buffart LM. Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses. BMC Res Notes 2019;12(01):164 [PubMed: 30902064]
9. Firnkorn D, Ganzinger M, Muley T, Thomas M, Knaup PConstruction and Application for the Lung Cancer Phenotype Database of the German Center for Lung Research. A generic data harmonization process for cross-linked research and network interaction. Methods Inf Med 2015;54(05):455–460 [PubMed: 26394900]
10. Bauer CRKD, Ganslandt T, Baum B, et al. Integrated data repository toolkit (IDRT). A suite of programs to facilitate health analytics on heterogeneous medical data. Methods Inf Med 2016;55(02):125–135 [PubMed: 26534843]
11. PCORnet. The national patient-centered clinical research network. Accessed April 20, 2021 at: https://pcornet.org/data/
12. OMOP Common Data Model – OHDSI. Accessed April 20, 2021 at: https://www.ohdsi.org/data-standardization/the-common-data-model/
13. i2b2: Informatics for integrating biology & the bedside. Accessed April 20, 2021 at: https://www.i2b2.org/about/intro.html
14. Boussadi A, Zapletal E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak 2017;17(01):120 [PubMed: 28806953]
15. CDISC. Clear Data. Clear Impact. Accessed April 20, 2021 at: https://www.cdisc.org/
16. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med 2009;48(05):408–413 [PubMed: 19621114]
17. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: a methodological review. J Biomed Inform 2016;60:352–362 [PubMed: 26944737]
18. 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). J Biomed Inform 2015;57:88–99 [PubMed: 26188274]
19. Bruland P, Breil B, Fritz F, Dugas M. Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM. Stud Health Technol Inform 2012;180:564–568 [PubMed: 22874254]

20. Navarro G. A guided tour to approximate string matching. ACM Comput Surv 2001;33(01):31–88

21. Nachimuthu SK, Lau LM. Applying hybrid algorithms for text matching to automated biomedical vocabulary mapping. AMIA Annu Symp Proc 2005;2005:555–559 [PubMed: 16779101]

22. MedDRA. Welcome to MedDRA. Accessed April 20, 2021 at: https://www.meddra.org/

23. McLaughlin VV, Benza RL, Rubin LJ, et al. Addition of inhaled treprostinil to oral therapy for pulmonary arterial hypertension: a randomized controlled clinical trial. J Am Coll Cardiol 2010;55(18):1915–1922 [PubMed: 20430262]

24. GALIE N. Ambrisentan in Pulmonary Arterial Hypertension. Randomized, double-blind, placebo-controlled, multicenter, efficacy studies (ARIES) group. Circulation 2008;117:2966–2968 [PubMed: 18541751]

25. Rubin LJ, Badesch DB, Barst RJ, et al. Bosentan therapy for pulmonary arterial hypertension. N Engl J Med 2002;346(12):896–903 [PubMed: 11907289]

26. Simonneau G, Barst RJ, Galie N, et al. ; Treprostinil Study Group. Continuous subcutaneous infusion of treprostinil, a prostacyclin analogue, in patients with pulmonary arterial hypertension: a double-blind, randomized, placebo-controlled trial. Am J Respir Crit Care Med 2002;165(06):800–804 [PubMed: 11897647]

27. Channick RN, Simonneau G, Sitbon O, et al. Effects of the dual endothelin-receptor antagonist bosentan in patients with pulmonary hypertension: a randomised placebo-controlled study. Lancet 2001;358(9288):1119–1123 [PubMed: 11597664]

28. Jing Z-C, Parikh K, Pulido T, et al. Efficacy and safety of oral treprostinil monotherapy for the treatment of pulmonary arterial hypertension: a randomized, controlled trial. Circulation 2013;127(05):624–633 [PubMed: 23307827]

29. Olschewski H, Simonneau G, Galiè N, et al. ; Aerosolized Iloprost Randomized Study Group. Inhaled iloprost for severe pulmonary hypertension. N Engl J Med 2002;347(05):322–329 [PubMed: 12151469]

30. Galiè N, Barberà JA, Frost AE, et al. ; AMBITION Investigators. Initial use of ambrisentan plus tadalafil in pulmonary arterial hypertension. N Engl J Med 2015;373(09):834–844 [PubMed: 26308684]

31. Oudiz RJ, Galiè N, Olschewski H, et al. ; ARIES Study Group. Long-term ambrisentan therapy for the treatment of pulmonary arterial hypertension. J Am Coll Cardiol 2009;54(21):1971–1981 [PubMed: 19909879]

32. Olschewski H, Hoeper MM, Behr J, et al. Long-term therapy with inhaled iloprost in patients with pulmonary hypertension. Respir Med 2010;104(05):731–740 [PubMed: 20153158]

33. Rubin LJ, Badesch DB, Fleming TR, et al. ; SUPER-2 Study Group. Long-term treatment with sildenafil citrate in pulmonary arterial hypertension: the SUPER-2 study. Chest 2011;140(05):1274–1283 [PubMed: 21546436]

34. Pulido T, Adzerikho I, Channick RN, et al. ; SERAPHIN Investigators. Macitentan and morbidity and mortality in pulmonary arterial hypertension. N Engl J Med 2013;369(09):809–818 [PubMed: 23984728]

35. Tapson VF, Torres F, Kermeen F, et al. Oral treprostinil for the treatment of pulmonary arterial hypertension in patients on background endothelin receptor antagonist and/or phosphodiesterase type 5 inhibitor therapy (the FREEDOM-C study): a randomized controlled trial. Chest 2012;142(06):1383–1390 [PubMed: 22628490]

36. Tapson VF, Jing Z-C, Xu K-F, et al. ; FREEDOM-C2 Study Team. Oral treprostinil for the treatment of pulmonary arterial hypertension in patients receiving background endothelin receptor antagonist and phosphodiesterase type 5 inhibitor therapy (the FREEDOM-C2 study): a randomized controlled trial. Chest 2013;144(03):952–958 [PubMed: 23669822]

37. Ghofrani H-A, D'Armini AM, Grimminger F, et al. ; CHEST-1 Study Group. Riociguat for the treatment of chronic thromboembolic pulmonary hypertension. N Engl J Med 2013;369(04):319–329 [PubMed: 23883377]

38. Ghofrani H-A, Galiè N, Grimminger F, et al. ; PATENT-1 Study Group. Riociguat for the treatment of pulmonary arterial hypertension. N Engl J Med 2013;369(04):330–340 [PubMed: 23883378]

39. Sandoval J, Torbicki A, Souza R, et al. ; STRIDE-4 investigators. Safety and efficacy of sitaxsentan 50 and 100 mg in patients with pulmonary arterial hypertension. Pulm Pharmacol Ther 2012;25(01):33–39 [PubMed: 22079088]

40. Sitbon O, Channick R, Chin KM, et al. ; GRIPHON Investigators. Selexipag for the treatment of pulmonary arterial hypertension. N Engl J Med 2015;373(26):2522–2533 [PubMed: 26699168]

41. Galiè N, Ghofrani HA, Torbicki A, et al. ; Sildenafil Use in Pulmonary Arterial Hypertension (SUPER) Study Group. Sildenafil citrate therapy for pulmonary arterial hypertension. N Engl J Med 2005;353(20):2148–2157 [PubMed: 16291984]

42. Benza RL, Barst RJ, Galie N, et al. Sitaxsentan for the treatment of pulmonary arterial hypertension: a 1-year, prospective, open-label observation of outcome and survival. Chest 2008;134(04):775–782 [PubMed: 18625676]

43. Barst RJ, Langleben D, Frost A, et al. ; STRIDE-1 Study Group. Sitaxsentan therapy for pulmonary arterial hypertension. Am J Respir Crit Care Med 2004;169(04):441–447 [PubMed: 14630619]

44. Oudiz RJ, Brundage BH, Galiè N, et al. ; PHIRST Study Group. Tadalafil for the treatment of pulmonary arterial hypertension: a double-blind 52-week uncontrolled extension study. J Am Coll Cardiol 2012;60(08):768–774 [PubMed: 22818063]

45. Galiè N, Brundage BH, Ghofrani HA, et al. ; Pulmonary Arterial Hypertension and Response to Tadalafil (PHIRST) Study Group. Tadalafil therapy for pulmonary arterial hypertension. Circulation 2009;119(22):2894–2903 [PubMed: 19470885]

46. Barst RJ, Langleben D, Badesch D, et al. STRIDE-2 Study Group. Treatment of pulmonary arterial hypertension with the selective endothelin-a receptor antagonist sitaxsentan. J Am Coll Cardiol 2006;47(10):2049–2056 [PubMed: 16697324]

47. Dootson A. Tracing data elements through a standard data flow. Pharmaceutical Programming 2011;41–2:59–69

48. Holmes D, McCabe MC. Improving precision and recall for Soundex retrieval. Paper presented at: Proceedings International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada; April 8–10, 2002:22–26

49. Cohen A. Fuzzywuzzy: Fuzzy string matching in Python. Accessed April 20, 2021 at: https://github.com/seatgeek/fuzzywuzzy

Table 1

Summary LLT mapping counts for all MH and AE data records

Instances	*n* (%) MH	*n* (%) AE
Total	37,105 (100)	58,170 (100)
Exact matches	27,777 (74.9)	49,910 (85.5)
Fuzzy matches	9,244 (24.9)	8,047 (13.8)
Unmapped	84 (0.2)	213 (0.37)

Abbreviations: AE, adverse event; LLT, lowest level term; MH, medical history.

Table 2

MH LLT mapping quality summary

MH map quality score	***n* (%)**
0: Exact match (data PRT and LLT)	5,375 (14.49)
1: Exact match (data PRT)	4,934 (13.3)
2: Exact match (data LLT)	15,286 (41.2)
3: Exact match (other)	2,182 (5.9)
4: Fuzzy match (high confidence)	9,210 (24.8)
5: Fuzzy match (medium confidence)	34 (0.1)
6: Unmapped	84 (0.2)

Abbreviations: LLT, lowest level term; MH, medical history; PRT, primary term.

Table 3

AE LLT mapping quality summary

AE map quality score	***n*** **(%)**
0: Exact match (data PRT and LLT)	18,414 (31.7)
1: Exact match (data PRT)	11,300 (19.4)
2: Exact match (data LLT)	14,434 (24.8)
3: Exact match (other 1)	4 (0.007)
4: Exact match (other 2)	5,758 (9.9)
5: Fuzzy match (high confidence)	6,824 (11.7)
6: Fuzzy match (medium confidence)	1,223 (2.1)
7: Unmapped	213 (0.37)

Abbreviations: AE, adverse event; LLT, lowest level term; PRT, primary term.

Table 4

MH and AE term hierarchy imputation summary

Level	Imputation quality score	*n* (%) MH	*n* (%) AE
HLT	1: No branches (deterministic)	24,161 (65.1)	36,202 (62.2)
HLT	2: Exact match	6,571 (17.7)	6,931 (11.9)
HLT	3: Fuzzy match	889 (2.4)	478 (0.8)
HLT	4: Default—first branch found	5,400 (14.6)	14,346 (24.7)
HLT	5: Unmapped	84 (0.2)	213 (0.37)
HLGT	1: No branches (deterministic)	36,578 (98.6)	56,535 (97.2)
HLGT	2: Exact match	284 (0.8)	659 (1.1)
HLGT	3: Fuzzy match	5 (0.01)	5 (0.009)
HLGT	4: Default—first branch found	154 (0.4)	758 (1.3)
HLGT	5: Unmapped	84 (0.2)	213 (0.37)
SOC	1: No branches (deterministic)	36,361 (98)	57,100 (98.2)
SOC	2: Exact match	417 (1.1)	122 (0.2)
SOC	3: Fuzzy match	143 (0.4)	2 (0.003)
SOC	4: Default—first branch found	100 (0.3)	733 (1.2)
SOC	5: Unmapped	84 (0.2)	213 (0.37)

Abbreviations: AE, adverse event; HLGT, high level group term; HLT, high level term; MH, medical history; SOC, system organ class.