HHS Public Access

Author manuscript

Clin Pharmacol Ther. Author manuscript; available in PMC 2020 April 01.

Published in final edited form as:

Clin Pharmacol Ther. 2020 April ; 107(4): 934–943. doi:10.1002/cpt.1787.

Development of a System for Post-marketing Population Pharmacokinetic and Pharmacodynamic Studies using Real-World Data from Electronic Health Records

Leena Choi[1], **Cole Beck**[1], **Elizabeth McNeer**[1], **Hannah L. Weeks**[1], **Michael L. Williams**[1], **Nathan T. James**[1], **Xinnan Niu**[2], **Bassel W. Abou-Khalil**[3], **Kelly A. Birdwell**[4], **Dan M. Roden**[2,4,5], **C. Michael Stein**[4,5], **Cosmin A. Bejan**[2], **Joshua C. Denny**[2,4], **Sara L. Van Driest**[4,6]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[3]Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[5]Department of Pharmacology, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[6]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

Abstract

Post-marketing population pharmacokinetic (PK) and pharmacodynamic (PD) studies can be useful to capture patient characteristics affecting PK or PD in real-world settings. These studies require longitudinally measured dose, outcomes, and covariates in large numbers of patients; however, prospective data collection is cost-prohibitive. Electronic health records (EHRs) can be an excellent source for such data, but there are challenges, including accurate ascertainment of drug dose. We developed a standardized system to prepare datasets from EHRs for population PK/PD studies. Our system handles a variety of tasks involving data extraction from clinical text using a natural language processing algorithm, data processing, and data building. Applying this system, we performed a fentanyl population PK analysis, resulting in comparable parameter estimates to a prior study. This new system makes the EHR data extraction and preparation process more efficient and accurate, and provides a powerful tool to facilitate post-marketing population PK/PD studies using information available in EHRs.

Corresponding Author: Leena Choi, Ph.D., R.Ph., Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Ave. Suite 1100, Nashville, TN 37203, Phone: 615-343-3497, leena.choi@vumc.org.

Author Contributions

L.C., E.M., H.L.W., N.T.J., D.M.R., C.M.S., J.C.D., and S.L.V.D wrote manuscript. L.C. designed research. L.C., C.B., E.M., H.L.W., M.L.W., N.T.J., X.N., B.W.A.K., K.A.B., C.A.B., J.C.D., and S.L.V.D performed research. L.C., C.B., and E.M. analyzed data. All authors approved final manuscript.

Conflict of Interest

Sara L. Van Driest is an associate editor for the journal Clinical Pharmacology and Therapeutics. All other authors declared no competing interests for this work.

Keywords

post-marketing population pharmacokinetic and pharmacodynamic study; medication dose extraction algorithm; data preparation; data processing; natural language processing; electronic health records; real-world data

Introduction

Pharmacokinetic (PK) and pharmacodynamic (PD) studies play an important role during all phases of drug development. In early phases, PK/PD studies inform development of drug dose and schedule. In later phases, PK/PD studies define dose adjustment for subpopulations (e.g., organ disfunction, genotype, drug-drug interactions). Population PK/PD studies in the post-marketing phase have the potential to capture patient characteristics affecting PK or PD in patients treated in real-world settings, where patients are more heterogeneous and diverse than participants in phase I-III trials who must meet strict inclusion/exclusion criteria.[1,2]

Population PK/PD analyses require longitudinally measured information on dose, outcomes, and potential covariates in a large number of patients. Population PK/PD modeling approaches can fit sparse data by taking a mechanistic modeling approach and borrowing information across a large number of subjects.[3,4] Until recently, prospective collection of such data has been cost-prohibitive, but now electronic health records (EHRs) can be an excellent source for such data. Real-world data (RWD) captured in the EHR present a unique opportunity to advance knowledge in this field. Data on drug doses, demographics, clinical covariates such as concomitant diseases and concomitant drug exposures, and clinically relevant outcomes are routinely documented in the EHR as part of clinical practice. Furthermore, important covariates already collected in the EHR are generally needed for the development of real-time clinical decision support systems.

There are multiple challenges to using RWD from EHRs in population PK/PD applications.[5,6] Data quality is one, and another is the need for automated data abstraction. Many post-marketing population PK/PD studies performed using EHRs to date have used manual curation methods that are not easily scalable,[7–9] and thus will not be useful for high-throughput and transparent high quality data abstraction. With “Big Data” sources such as EHRs, data extraction and processing may be error prone, tedious and time consuming. Validated programs to perform each of many steps are required.

The goal of this study was to develop a standardized and efficient system for data extraction and preparation from EHRs for population PK/PD studies, which could be generalized beyond the specific drugs studied. To this end, we standardized the entire data preparation procedure from extraction from the EHR to PK/PD data building. We present results for four test drugs: tacrolimus, lamotrigine, fentanyl, and dexmedetomidine.

Methods

Study Design and Data Source

This study was approved by the Vanderbilt Institutional Review Board. The key data elements required to perform population PK/PD studies include medication dose, drug concentration levels and/or phenotype for drug response, and subject characteristics such as demographics, laboratory and genotype data. To develop a system for generating datasets for post-marketing PK/PD studies using EHRs, we selected two medications as test drugs for each of the two most common routes of drug administration: oral (tacrolimus and lamotrigine) and IV (fentanyl and dexmedetomidine) administration.

For the tacrolimus and lamotrigine cohorts, we extracted data from the Synthetic Derivative (SD) at VUMC, a de-identified database of clinical records derived from Vanderbilt's EHR system. Tacrolimus was selected since a manually-curated gold standard validated dose dataset was available from a previous study (n = 446, further details below).[10] Lamotrigine was selected since the dosing regimens for this drug are complex and very different from those of tacrolimus, and drug concentration levels are measured as part of routine therapeutic drug monitoring. For lamotrigine, we created a new cohort as follows. We first identified SD records containing the keywords 'lamotrigine' or 'Lamictal,' which yielded 13,176 subjects. We then selected subjects whose ages were between 18 to 70 when their first lamotrigine level measurement was available and who had an ICD-9-CM or ICD-10-CM (The International Classification of Diseases, Ninth and Tenth Revision, Clinical Modification) billing code for epilepsy. This yielded 2,475 subjects; of those, 305 subjects who had at least 3 lamotrigine levels and 3 dose information within 5 years of data remained as the final cohort. For each subject in the cohort for these two medications, we identified all clinical notes generated on the same date as when a drug concentration level was available, and from these notes medication dosing information was extracted. For each cohort, other structured data required to perform PK/PD studies were also extracted from the SD.

In contrast to oral medications, exposures to IV medications typically happen during an inpatient admission, and all administered doses are documented as structured data in the EHR. For IV medications, we selected fentanyl (n = 498) and dexmedetomidine (n = 411) as we have been collaborating with investigators of other ongoing studies with these medications given to pediatric populations,[11,12] which provided us easier access to the data and also helped us to validate the dose data and programs as needed. In addition, fentanyl IV dose data include both bolus and infusion while dexmedetomidine data contain mainly infusion, allowing us to work with diverse IV dosing patterns. The raw dose data for fentanyl and dexmedetomidine were extracted from Vanderbilt Enterprise Data Warehouse.[11] Additionally, other data required to perform PK/PD studies were obtained from the Vanderbilt EHRs for each cohort.

Data Extraction

Data in EHRs can be largely divided into two types: 1) structured data that are readily available as various tables and 2) unstructured data that need to be extracted from free text sources such as clinical notes. Drug levels, laboratory values, and demographic information

are structured data. These were extracted from the EHR for each medication cohort using standard structured query language (SQL). Fentanyl and dexmedetomidine dose information were also extracted using SQL from electronic medication administration records (MAR) and flowsheet data in the EHR. We extracted tacrolimus and lamotrigine dose data from two data sources: 1) RxStar, which is a structured database for e-prescriptions, and 2) clinical notes, which have unstructured data. To extract dose information from clinical notes, we used and compared several existing natural language processing (NLP) systems that were developed for medication and signature information extraction: MedEx[13] that was originally implemented in the SD, CLAMP[14] that incorporates an extended version of MedEx, and MedXN,[15] which has been reported as outperforming MedEx, as well as our own NLP system, medExtractR[16] that is a specialized NLP system for PK/PD studies.

A Tool for Gold Standard Data Generation and Gold Standard Datasets

To evaluate dose information extracted from RxStar and clinical notes for the oral medications, lamotrigine and tacrolimus, we compared the extracted data to gold standard datasets that were manually curated by clinical experts. For tacrolimus, we used a clinically validated tacrolimus dose data from a previous study,[10] which contains 4,150 records from 446 subjects. For lamotrigine, we generated a new gold standard dataset. To assist with the manual annotation of the data by the clinical experts, we developed a tool using the R[17] programming language (Supplemental Material Text 2). Using this tool, we generated the lamotrigine dose gold standard dataset of 4,401 records for 305 subjects.

Development of a System for Post-marketing PK/PD Studies

With the extracted raw data files, we developed programs for processing raw data and building PK/PD datasets using the R[17] programming language. For each extracted data type, we wrote scripts to perform post-processing. As examples of tasks, the scripts remove invalid data, impute missing data by pre-defined rules, convert all time variables from different sources to a consistent form, and match several variables with time at which they were administered or measured, so that they can be merged appropriately to build the final PK/PD datasets. For the data building, we also wrote scripts to build datasets for PK/PD modeling with oral and IV administration separately, which are analyzable using NONMEM.[18] The data building programs were developed through an iterative procedure – that is the programs were continuously updated until the output was correct. As this algorithm was completely rule-based, there was no uncertainty in when the correct result was produced. Nevertheless, a sample of the output was compared to manually constructed PK/PD data.

There are many steps in data processing and building with varying degrees of complexity. As described above, we first wrote a drug-specific R script to perform diverse tasks from importing several raw data files through building PK/PD datasets for each of four medications, two oral and two IV medications. As some steps are very specific to data type and medications, while others are common, we modularized the major steps, which helped standardize the entire data preparation. The major procedures included data extraction, data processing, and data building. Under each procedure, we created modules depending on the data element, task to perform, and type of PK/PD models.

Although the process was the same, the original script was not modular, and hence difficult to maintain and understand. We refactored this code into various modules, or R functions, and collected them in an R package. We created modules using subsets of the original lengthy R scripts that can read each data file, call in relevant R functions to perform each job, and generate output data. When developing R functions and writing the scripts, two programmers worked together to validate the programs. After the development was completed, we selected relevant modules to generate the final PK dataset for each medication and checked whether this output matched the original output generated using the drug-specific script. The process was iterated until the system yielded the same output as the original non-modularized script.

Validation of Extracted Dose Data

We evaluated whether medication dose information obtained from e-prescriptions (i.e., RxStar dose) would provide sufficient data for PK/PD studies. For each validated dose record in the gold standard dataset, we identified the RxStar dose at the closest time point before that record and calculated the percent of matched doses out of the total number of validated doses in the gold standard dataset for both tacrolimus and lamotrigine. In the event multiple RxStar dose values were associated with the same validated dose, we considered them to be matched if at least one of them matched the validated dose.

We also evaluated the extracted dose from clinical notes using different NLP systems. For each validated dose in the gold standard dataset, we determined whether the same dose was found in the extracted dose on the same date by each of four NLP systems (i.e., MedEx, CLAMP, MedXN and medExtractR). As multiple doses can be extracted from clinical notes on the same date (due to multiple mentions of drug dose information on the same date), if the same dose was found in any of an NLP extracted doses, it was considered matched with the gold standard dose, which would provide the best possible performance. We calculated the percent matched dose out of the number of doses in the gold standard dataset for both medications.

A Case Study

We performed a case study using fentanyl that represents a medication with relatively complex IV infusion/bolus administration. Results of the current analysis of 498 subjects are compared to previously reported a fentanyl population PK study with a smaller cohort of 130 subjects.[11]

Results

A System for Data Generation for PK/PD Studies using EHRs

The overall system for generating datasets for post-marketing PK/PD studies using EHRs is presented schematically in Figure 1. The system was divided into three major procedures: Data Extraction, Data Processing, and Data Building, each consisting of multiple modules. Once the type of study was determined, modules were selected based on data availability as described below.

Data Extraction Procedure

Extraction of structured data (e.g., laboratory values) from EHRs is generally straightforward. In contrast, accurate extraction of information from clinical notes (e.g., drug dose regimens) by NLP or machine learning methods is a major challenge. Our module for oral dose extraction is presented below. A generalized module for extracting clinical conditions, Extract-Phenotype, is under development.

Medication dose extraction module: Extract-Med extracts dose information from clinical text using an NLP system we developed, medExtractR. More details of its usage can be found in Weeks *et al.*[16] The main inputs to medExtractR are a clinical note and a list of drug names of interest for which dosing information should be extracted. The drug names can include variations such as the generic name, brand name, and abbreviations. This module returns an output of dose data that includes drug name, frequency, intake time, strength and dose amount. Keywords such as ‘increase’ or ‘change to’ which indicate whether a dose regimen is current (called ‘dose change’) are extracted when present. The module is also able to extract the time of the last dose when present in the text. The major steps of this module are illustrated in Figure 2. Depending on the drug of interest, some default parameters of medExtractR may need to be adjusted, such as the unit of the drug (e.g., ‘mg’) and degree of misspellings allowed for drug names (known as edit distance). Dictionaries for entities such as frequency or intake time may also need to be updated with any institution-specific phrases. After specifying all parameters, the algorithm can be run and will return a table listing the entity name, extracted expression, and the start/stop positions within the note, which is an input for a data processing module, Pro-Med-NLP.

Post-Extraction Data Processing Procedure

The output data from the extraction modules as well as all raw data directly extracted from the EHR are not usable until they have been mapped to a structured analytical data. We envisioned six modules for processing six different types of data extracted from the EHR (see Data Processing in Figure 1). Two modules for medication dose processing, the most challenging task, were developed separately for NLP-extracted oral dose data and structured dose data, named Pro-Med-NLP and Pro-Med-Str, as their data processing tasks differ widely.

NLP-extracted medication dose processing module: Pro-Med-NLP can process output from medExtractR[16] or three other NLP systems for medication information extraction: MedEx,[13] CLAMP,[14] and MedXN[15] (Figure 3). As it was challenging to process the raw extracted data, especially for drugs prescribed multiple times a day, we developed a rigorous post-processing algorithm that was implemented in Pro-Med-NLP. The details of this algorithm can be found in McNeer *et al.*[19] Briefly, step 1 (Parsing) transforms each NLP output to a standardized form through its parse function written specific to each NLP system. After this step, the next steps are common across all NLP systems, allowing greater generalizability, which include: step 2 (Pairing) uses regular expressions to match drug names of interest (e.g., “lamot|lamictal|ltg” for lamotrigine), and is able to handle special cases such as missing frequency; and step 3 (Building) removes redundancies, and calculates dose intake and daily dose (Figure 3).

Structured medication dose processing module: Pro-Med-Str processes structured medication dose data, consisting of two parts: intravenous (IV) infusion/bolus data and e-prescription data (Figure 1). While processing structured medication data obtained from e-prescriptions (**Part II**) is straightforward with relatively simple data cleaning steps such as making numeric variables for strength, frequency and dose, and removing duplicates, the data processing for IV infusion/bolus dose (**Part I**) presents a different challenge compared to NLP-extracted dose data. In our EHR, IV infusion drug data are obtained from two different data sources: electronic medication administration records (MAR) data and flowsheet data. These were processed separately as they are in different forms. In brief, the part I of the module processes MAR data by sub-setting data to the drug of interest using a medication name list [e.g., Fentanyl (diluted), Fentanyl injection], extracting dose information, and splitting infusion and bolus data. It also removes irrelevant data (e.g., pharmacy dispensing data). Next, it processes the flowsheet data, which involves multiple cleaning steps as illustrated in Figure S1. In the dexmedetomidine dataset, for example, removal of duplicates reduced the number of rows from 104,222 to 98,217; further removal of erroneous data (e.g., missing rate) resulted in 92,992 rows in the dataset.

Other data processing modules: Pro-Drug Level, Pro-Laboratory, and Pro-Demographic modules process structured drug level, laboratory data, and demographic data, respectively. Common functionality shared by all modules includes a step to standardize the date/time variable to enable merging of several data elements across different files, and a data checking step based on investigator input (e.g., record errors, irrelevant data, outliers, and missing data). Aspects unique to each module are summarized in the box under each module in Figure 1. Details of these modules can be found in Supplementary Material Text 1.

PK/PD Data Building Procedure

The data building modules include Build-PK-Oral, Build-PK/PD-Oral, Build-PK-IV, and Build-PK/PD-IV. Currently, the modules build datasets to perform PK/PD analyses using NONMEM,[18] the most commonly used software for population PK/PD analysis. Example studies that can be performed with the developed system are presented at the bottom of Figure 1.

Evaluation of Extracted Dose Data

Oral dose data can be obtained from both a structured database for e-prescriptions such as RxStar and an unstructured data source such as clinical notes using NLPs. Compared to the validated dose in gold standard datasets, Table 1 presents the evaluation results on tacrolimus and lamotrigine dose data extracted from the structured data (RxStar) and unstructured data by different NLP systems (MedEx,[13] CLAMP,[14] MedXN,[15] and medExtractR).[16] As RxStar was implemented in Vanderbilt University Medical Center (VUMC) as an e-prescription database in 2003, the evaluation of RxStar dose data was restricted to validated doses in the gold standard datasets from January 1, 2004, yielding 3,631 for tacrolimus (out of 4,144 total validated daily dose records) and 2,083 for lamotrigine (out of 2,110 total validated daily dose records) as the denominator for the evaluation. RxStar data were missing dose data in 45.2% and 24.4% of validated dose

records for tacrolimus and lamotrigine, respectively. Restricted to the gold standard daily doses found in RxStar (i.e., 1,990 and 1,574 valid dose records, which are 54.8% and 75.6% for tacrolimus and lamotrigine validated dose records, respectively), 47.8% and 66.6% of records matched for tacrolimus and lamotrigine, respectively. Overall, RxStar data matched 26.2% and 50.4% of the validated doses for tacrolimus and lamotrigine, respectively. The low agreement is likely due to frequent dose changes for these medications during therapeutic drug monitoring. The evaluation of medication extraction from unstructured data showed that medExtractR outperformed the other 3 NLP systems with agreement of 95.0% and 93.1% for tacrolimus and lamotrigine, respectively.

Case Study

Using the system architecture illustrated in Figure 1, we successfully generated a dataset for fentanyl population PK study with a cohort of 498 subjects, which includes 118 subjects used in a previously reported fentanyl population PK study,[11] some of which were excluded by our inclusion/exclusion criteria. The number of bolus and infusion doses and samples per subject were similar between the two cohorts. The number of bolus doses were the median 10 (interquartile range, IQR, 7 – 14) for the prior cohort, and the median 7 (IQR 5 – 10) for the current cohort, while the median and IQR for the number of infusion doses were the same for both cohorts (the median 3, IQR 2 – 5). The number of samples per subject for the prior and current cohorts were the median 6 (IQR 4 – 10) and the median 5 (IQR 3 – 7), respectively.

After all raw data files were extracted from the Vanderbilt EHRs, the final dataset was generated using the following modules: Pro-Med-Str, Pro-Drug Level, Pro-Laboratory, Pro-Demographic, and Build-PK-IV. When we conducted the previous study with 130 subjects, [11] it took more than one year to generate the final dataset for the analysis, while it took only about two weeks to prepare the final dataset with our new system. The major time-determining step in using the system was data checking by the investigator for potential errors identified by the system. After the data checking was completed, a final dataset was generated in 10 minutes. With these data, we performed population PK analysis using the same model reported in the previous study. As presented in Table 2, the major PK parameter estimates from both cohorts were comparable (for subjects with 70 kg, total clearance, CL: 49 vs. 45 L/hr; central volume of distribution, V_1: 200 vs. 230 L; intercompartmental clearance, Q: 35 vs. 22 L/hr; peripheral volume of distribution, V_2: 520 vs. 580 L). The observed differences in PK parameter estimates between the prior and current studies may be due to differences in subjects' characteristics.

Discussion

We developed a system using EHRs to generate data suitable for post-marketing PK/PD studies across diverse medications. The case study showed that our system significantly improved the efficiency of data building and standardized the entire data generation procedure. Using the generated dataset, we successfully performed a fentanyl population PK analysis, with much faster data preparation than the prior manual process.

Using a modular system provides several benefits. First, this helped us to validate the programs and make them more generalizable. Second, because modules work as building blocks in the system, only necessary modules need to be used to build the final dataset depending on type of data and study. Third, modules can be more easily customized if needed. Each module can be improved in an independent manner and the improved module can be easily integrated back into the system. Thus, the modularization facilitated standardizing the entire data preparation procedure, which may otherwise be difficult to achieve for such a complicated process involving many steps. Fourth, all modules were written in R, a freely available and widely used statistical software. We plan to make our modules publicly available by early spring 2020. The code can be easily adapted to other EHR systems, helping to ensure the developed system will be generalizable and scalable.

Our study confirmed that structured medication dose data extracted from an e-prescription database such as RxStar was inadequate for analyses. Although it is easy to obtain the structured e-prescription dose data, more complete data are required for medication-related studies including population PK/PD studies. The discrepancy between validated dose and the e-prescription records from the VUMC specific e-prescribing tool, RxStar, may not be true for other EHR e-prescribing tools. A discrepancy between actual dose and initially prescribed dose may be more likely for routinely monitored and adjusted medications such as those used in this study compared to many other common medications. Of note, we analyzed fentanyl and dexmedetomidine using inpatient exposures to the drugs via IV route. We were thus able to analyze dose data from documented inpatient administrations, which are structured data obtained from inpatient flowsheet and MAR databases. Omitted or inaccurate dose data may be present in these sources. As a long sequence of IV dose data are usually available for each patient, these inpatient IV dose data are sufficiently accurate as a dosing history.

Our study also showed that the performance of existing NLP systems for medication dose extraction varied widely from 64.1% to 92.3%, supporting the need for further improvement. High performance is especially important to those NLP systems designed for medication-related studies, including PK/PD studies. Our more specialized algorithm medExtractR outperformed all other tested NLP systems including MedEx[13] that was deployed as a general-purpose medication extraction tool in our EHR. Our performance metrics may be inflated, as multiple doses can be extracted on the same date and any one match was considered success. Thus, further methodological development for identifying correct dose among multiple extracted doses or accounting for incorrect dose is warranted. On the other hand, as dose extracted from EHRs is never perfect, further investigation on how inaccurate dose data affect the results of PK/PD analysis is the focus of ongoing study.

In addition to dose retrieval, other important information such as clinical outcomes may need to be abstracted from clinical notes, which may be a covariate or serve as an outcome for PD studies. Although there is great potential, population PD studies have been underutilized. Use of EHR data may facilitate PD analyses as clinical observations for drug effects or adverse events are often documented as part of clinical care (e.g., recurrent cardiac events,[20] ACE inhibitor induced cough,[21] vancomycin-induced drug reaction).[22] However, phenotyping can impose different challenges depending on how it is defined. A sophisticated

phenotyping algorithm may need to be developed to abstract PD measures, and further development of NLP may be necessary in order to utilize EHR data for more diverse population PD studies. Examples from prior work include phenotype algorithms such as PheKB[23] and PheWAS[24] codes. Along this line of research, Extract-Phenotype and Pro-Phenotype are currently under development.

Patient characteristics identified from a post-marketing population PK/PD study can be utilized to find a better dose for individual patients. One important patient characteristic that can affect PK/PD profile is genotype. BioVU,[10] the DNA biobank at VUMC, couples genotypes to de-identified EHRs. Using this data, our system can facilitate pharmacogenomic studies. Specific pharmacogenomic modules are in development. Patient characteristics impacting drug response in a clinically meaningful way, including genotype, laboratory results, or other factors, can be included in population PK/PD analyses in order to define the impact of these factors on dose. Clinical implementation of this approach can be facilitated by Bayesian population PK/PD modeling, as each model could be seamlessly updated with our new system after each patient's data are entered (i.e., using the posterior distribution as a prior for a new dataset), and the updated model can provide an optimal dose for that individual patient. Bayesian population PK/PD models would provide a cohesive framework to incorporate all available patient information including uncertainty in PK/PD parameter estimates for individual patients. In this approach, the optimal dose for each patient is based on a hierarchical model which posits that the effect of patient characteristics on PK profile or drug response are similar, but not identical, across individuals. Optimal dose prediction for a patient with sparse or inaccurately measured data is strongly informed by other patients in the population, while predictions for a patient with frequent, accurate measurements rely more heavily on that patient's data.

This study has several limitations. First, the system was developed using four test medications. The two oral medications, tacrolimus and lamotrigine, were selected as representative of simple to complex prescription patterns. Although our system is expected to handle a wide range of medications, validation with other medications is required. Second, our system was developed using data from the VUMC EHR, and requires further validation and potentially customization for other EHR systems. There are also limitations inherent to the use of EHR data. For example, for oral medications, exact dosing time may not be available. If the time when the last dose was taken, which is the most important dosing time for PK analysis, is present in clinical text, our NLP module can extract this information. If not clinically documented, utility of population PK modeling using EHRs would be limited to medications with long half-life, where population PK modeling could provide a good approximation and identify patient characteristics affecting PK/PD profile.[25] EHR-based data collection is also limited to drugs, outcomes, and covariates that are routinely documented as part of clinical care, precluding analysis of over-the-counter medications as target drugs for PK analysis or as covariates (e.g., as inducers or inhibitors of drug metabolism enzymes).

Currently, a systematic approach of using EHRs in post-marketing population PK/PD studies is lacking. We developed a system based on standardized modules, making the process more efficient, less error prone, and more transparent, and ultimately enhancing

reproducible research. We hope that our system will provide potential solutions to overcome some challenges discussed by many authors, and can be a cornerstone for further development of a standardized data building and processing pipeline. This system could be sharable within and across institutions, and would ultimately pave a way to facilitate post-marketing population PK/PD studies using EHRs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This study is funded by National Institutes of Health (NIH) GM124109, GM 131770, and Doris Duke CSDA 2017075.

References

1. Balakrishnan AS et al. Minority Recruitment Trends in Phase III Prostate Cancer Clinical Trials (2003 to 2014): Progress and Critical Areas for Improvement. J. Urol 201, 259–267 (2019). [PubMed: 30218761]
2. Fisher JA & Kalbaugh CA Challenging assumptions about minority participation in US clinical research. Am J Public Health 101, 2217–2222 (2011). [PubMed: 22021285]
3. Steimer JL, Vozeh S, Racine-Poon A, Holford N & O'Neill R The Population Approach: Rationale, Methods, and Applications in Clinical Pharmacology and Drug Development In Pharmacokinetics of Drugs 110, 405–451 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1994).
4. Sheiner L & Wakefield J Population modelling in drug development. Statistical Methods in Medical Research 8, 183–193 (1999). [PubMed: 10636334]
5. Liu Q, Ramamoorthy A & Huang S-M Real-World Data and Clinical Pharmacology: A Regulatory Science Perspective. Clin. Pharmacol. Ther 106, 67–71 (2019). [PubMed: 30964944]
6. Van Driest SL & Choi L Real-World Data for Pediatric Pharmacometrics: Can We Upcycle Clinical Data for Research Use? Clin. Pharmacol. Ther 106, 84–86 (2019). [PubMed: 30942897]
7. Kakara M et al. Population pharmacodynamic analysis of LDL-cholesterol lowering effects by statins and co-medications based on electronic medical records. Br J Clin Pharmacol 78, 824–835 (2014). [PubMed: 24734885]
8. Hornik CP et al. Electronic Health Records and Pharmacokinetic Modeling to Assess the Relationship between Ampicillin Exposure and Seizure Risk in Neonates. J. Pediatr 178, 125–129.e1 (2016). [PubMed: 27522443]
9. Ku LC et al. Use of Therapeutic Drug Monitoring, Electronic Health Record Data, and Pharmacokinetic Modeling to Determine the Therapeutic Index of Phenytoin and Lamotrigine. Therapeutic Drug Monitoring 38, 728–737 (2016). [PubMed: 27764025]
10. Birdwell KA et al. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenetics and Genomics 22, 32–42 (2012). [PubMed: 22108237]
11. Van Driest SL et al. Pragmatic Pharmacology: Population Pharmacokinetic Analysis of Fentanyl using Remnant Samples from Children after Cardiac Surgery. Br J Clin Pharmacol (2016).doi:10.1111/bcp.12903
12. Shuplock JM et al. Association between perioperative dexmedetomidine and arrhythmias after surgery for congenital heart disease. Circ Arrhythm Electrophysiol 8, 643–650 (2015). [PubMed: 25878324]
13. Xu H et al. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 17, 19–24 (2010). [PubMed: 20064797]
14. Soysal E et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 25, 331–336 (2017).

15. Sohn S et al. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 21, 858–865 (2014). [PubMed: 24637954]
16. Weeks HL et al. medExtractR: A medication extraction algorithm for electronic health records using the R programming language. MedRxiv 19007286 (2019).doi:10.1101/19007286
17. R Core Team: A Language and Environment for Statistical Computing.
18. Beal S, Sheiner L & Boeckmann A NONMEM User's Guides. (Icon Development Solutions, Ellicott City, MD).
19. McNeer E, Beck C, Weeks HL, Williams ML & Choi L A post-processing algorithm for building longitudinal medication dose data from extracted medication information using natural language processing from electronic health records. bioRxiv 775015 (2019).doi:10.1101/775015
20. Delaney JT et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin. Pharmacol. Ther 91, 257–263 (2012). [PubMed: 22190063]
21. Mosley JD et al. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. Pharmacogenomics J 16, 231–237 (2016). [PubMed: 26169577]
22. Konvinse KC et al. HLA-A*32:01 is strongly associated with vancomycin-induced drug reaction with eosinophilia and systemic symptoms. J. Allergy Clin. Immunol 144, 183–192 (2019). [PubMed: 30776417]
23. Kirby JC et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 23, 1046–1052 (2016). [PubMed: 27026615]
24. Denny JC et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26, 1205–1210 (2010). [PubMed: 20335276]
25. Passey C et al. Dosing equation for tacrolimus using genetic variants and clinical factors. Br J Clin Pharmacol 72, 948–957 (2011). [PubMed: 21671989]

Study Highlights

What is the current knowledge on the topic?

Collection of data required to perform post-marketing population pharmacokinetic (PK) and pharmacodynamic (PD) studies has been cost-prohibitive. Many post-marketing PK/PD studies performed using electronic health records (EHRs) to date have used manual curation methods that are not easily scalable and may not provide transparent high quality data abstraction.

What question did this study address?

Can we prepare datasets from EHRs for post-marketing PK/PD studies in a more standardized, scalable, and efficient way?

What does this study add to our knowledge?

We developed a new system for data extraction and preparation from EHRs for post-marketing PK/PD studies that is standardized, scalable, and efficient.

How might this change clinical pharmacology or translational science?

The new system would facilitate post-marking PK/PD studies using EHR data by making the entire data preparation process more efficient, less error prone, and more transparent, and it would ultimately enhance reproducible research.

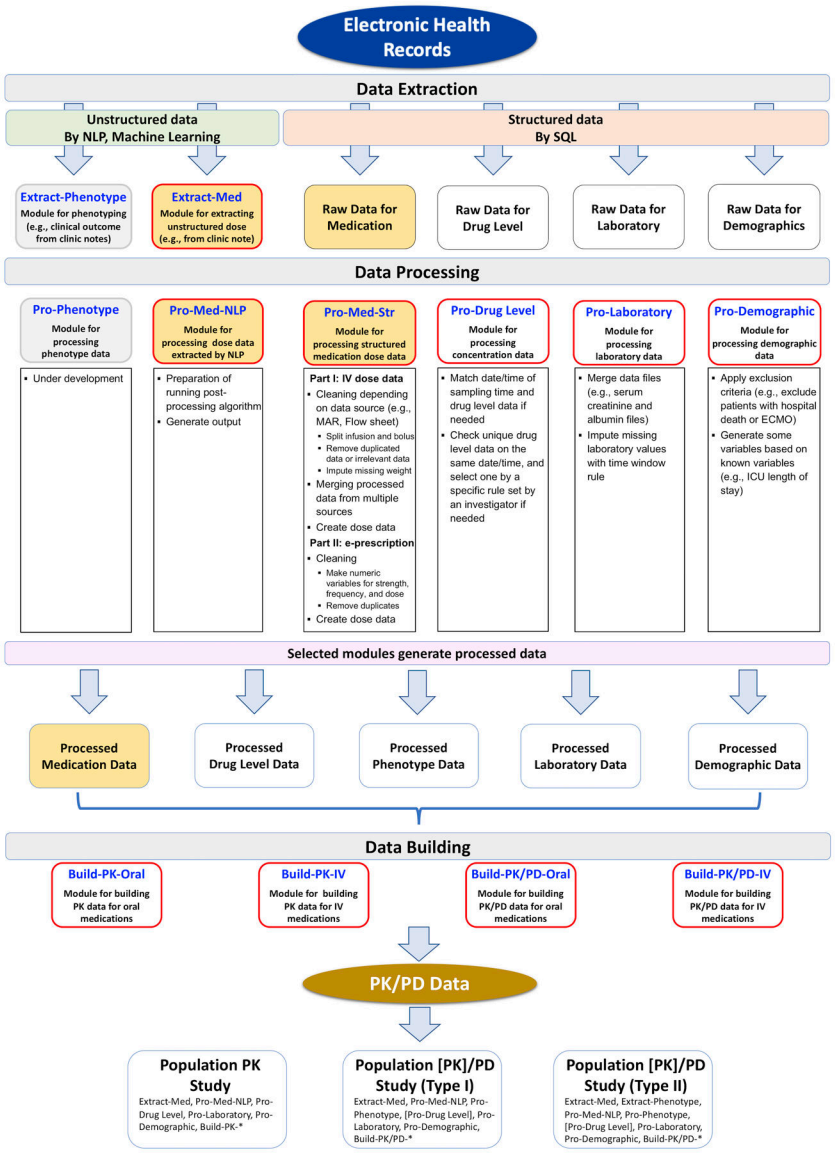

Figure 1.

Schematic presentation of a system for post-marketing PK/PD studies using EHRs. The system consists with three major procedures. “Extract-,” “Pro-,” and “Build-” are modules for Data Extraction, Data Processing, and Data Building procedures, respectively, and are represented by a red box or gray box (under development). As the system can handle two types of medication data (i.e., oral and intravenous route), medication related data and modules are represented in yellow boxes. The raw data extracted directly from the EHRs and the processed data are represented in black color font. Under Data Processing, the major tasks unique to each module are listed in each box, but common tasks across modules are not shown. A final PK/PD dataset generated after applying a data building module is shown in gold oblique circle. Example studies along with a list of modules that can be used to perform each study are also presented. Pro-Laboratory and Pro-Demographic are modules commonly included for most studies, while Pro-Drug Level and Pro-Phenotype should be

included for a study with PK and PD components. [PK]/PD in the example studies represents PK/PD or PD depending on a study design, where Pro-Drug Level should be included in the list of modules if PK component is included. * represents Oral or IV in “Build-PK-” or “Build-PK/PD-.”

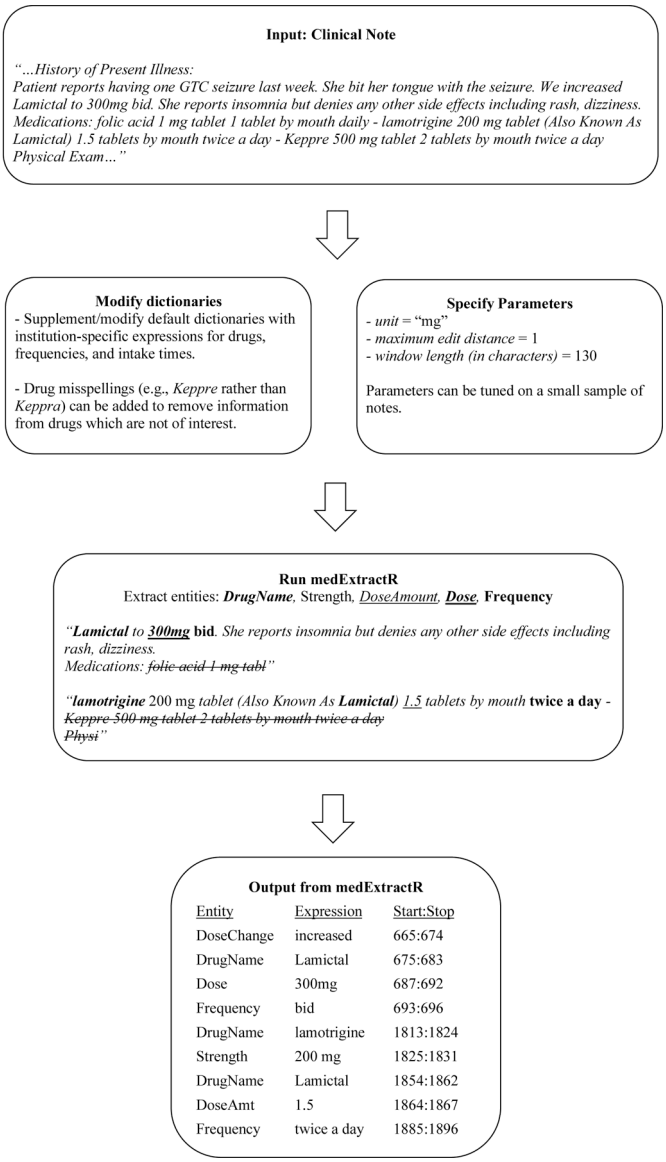

Figure 2.

Illustration of the medication dose extraction module, Extract-Med. This module extracts dose information from clinical text using a natural language processing algorithm, medExtractR, for which parameters are specified and dictionaries are modified depending on institution and medication. It returns an output of dose data that includes drug name, frequency, intake time, and either strength and dose amount or dose given intake.

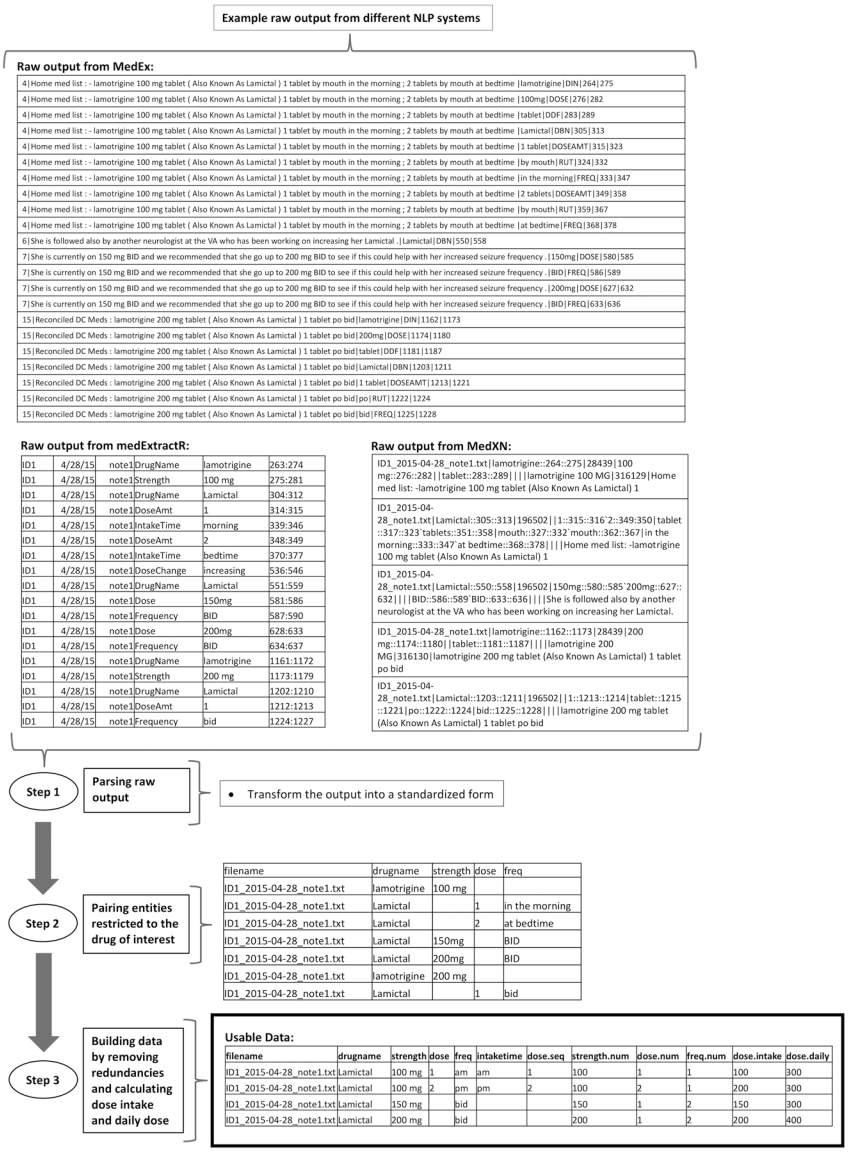

Figure 3.

Major steps of data processing module, Pro-Med-NLP, for medication dose information extracted by natural language processing (NLP) systems. The module can process raw output from four different NLP systems, MedEx, CLAMP, MedXN and medExtractR, of which examples of three NLP systems output are presented at the top. After transforming each output into a standardized form using parsing function specific to each NLP output (Step 1), next steps are common to all NLP outputs, which includes pairing associated entities (Step 2) and building a dataset in an analyzable format (Step 3).

Table 1

Evaluation of dose data extracted via multiple systems: an e-prescription database, RxStar, and several natural language processing (NLP) systems applied to clinical notes (MedEx, CLAMP, MedXN and medExtractR)

	e-prescriptions (RxStar)			Clinical notes (by NLP)			
	Missing[a]	Among doses found[b]	Overall[c]	MedEx	CLAMP	MedXN	medExtractR
Tacrolimus	45.2%	47.8%	26.2%	72.3%	65.9%	92.3%	95.0%
Lamotrigine	24.4%	66.6%	50.4%	78.5%	64.1%	89.4%	93.1%

a: % of validated dose records that were not found in RxStar;

b: % of matched dose records between the gold standard and RxStar out of those found in RxStar, for which the denominator is (1 – a)% [e.g., 1,990 (54.8%) tacrolimus and 1,574 (75.6%) lamotrigine validated dose records found in RxStar];

c: % of matched dose records between the gold standard and RxStar out of all validated doses, calculated by $c = 100 \times (1 - a) \times b$ [for example of tacrolimus, $26.2\% = 100 \times (1 - 0.452) \times 0.478$].

Table 2

PK model parameter estimates from fentanyl population PK studies

	Previous manually-reviewed study[14] (N = 130)	Current system-generated study (N = 498)
Parameters	**Estimates (SE)**	**Estimates (SE)**
$CL = \theta_1 (wgt/70)^{\theta 2}$		
θ_1	49.0 (8.2)	45.1 (3.7)
θ_2	0.9 (0.1)	0.8 (0.04)
$V_1 = \theta_3 (wgt/70)^{\theta 4}$		
θ_3	198.8 (107.8)	231.8 (45.5)
θ_4	0.7 (0.2)	0.7 (0.1)
$Q = \theta_5 (wgt/70)^{\theta 6}$		
θ_5	35.3 (17.0)	21.9 (4.3)
θ_6	1.3 (0.2)	1.0 (0.1)
$V_2 = \theta_7 (wgt/70)^{\theta 8}$		
θ_7	520.9 (181.3)	580.2 (112.8)
θ_8	1.2 (0.2)	1.1 (0.1)
ω^2_{CL} (%CV)	70 (13)	62 (5)
ω^2_{V1} (%CV)	64 (19)	71 (9)
$\sigma^2_{proportional}$ (%CV)	52 (5)	56 (2)
$\sigma^2_{additive}$ (ng mL^{-1})	0.01 (0.01)	0.002 (0.001)

SE, the standard error; CL, total clearance (L/hr); Q, intercompartmental clearance (L/hr); V_1, volume of distribution for the central compartment (L); V_2, volume of distribution for the peripheral compartment (L); CV, coefficient of variation; wgt, body weight in kg; ω^2CL and ω^2V1, the variance for η_iCL and η_iV1, respectively; $\sigma^2_{proportional}$ and $\sigma^2_{additive}$, the proportional and additive residual error variance, respectively.