ARTICLE

Applications of Advanced Natural Language Processing for Clinical Pharmacology

Joy C. Hsu[1,*,†], Michael Wu[2,†], Chloe Kim[2], Bianca Vora[1], Yi Ting (Kayla) Lien[1], Ashutosh Jindal[1], Kenta Yoshida[1], Sonoko Kawakatsu[1,3], Jeremy Gore[4], Jin Y. Jin[1], Christina Lu[2], Bingyuan Chen[2,‡] and Benjamin Wu[1,‡]

Natural language processing (NLP) is a branch of artificial intelligence, which combines computational linguistics, machine learning, and deep learning models to process human language. Although there is a surge in NLP usage across various industries in recent years, NLP has not been widely evaluated and utilized to support drug development. To demonstrate how advanced NLP can expedite the extraction and analyses of information to help address clinical pharmacology questions, inform clinical trial designs, and support drug development, three use cases are described in this article: (1) dose optimization strategy in oncology, (2) common covariates on pharmacokinetic (PK) parameters in oncology, and (3) physiologically-based PK (PBPK) analyses for regulatory review and product label. The NLP workflow includes (1) preparation of source files, (2) NLP model building, and (3) automation of data extraction. The Clinical Pharmacology and Biopharmaceutics Summary Basis of Approval (SBA) documents, US package inserts (USPI), and approval letters from the US Food and Drug Administration (FDA) were used as our source data. As demonstrated in the three example use cases, advanced NLP can expedite the extraction and analyses of large amounts of information from regulatory review documents to help address important clinical pharmacology questions. Although this has not been adopted widely, integrating advanced NLP into the clinical pharmacology workflow can increase efficiency in extracting impactful information to advance drug development.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Natural language processing (NLP) is commonly used and is embedded in various aspects of our lives. However, NLP has not been widely evaluated and utilized to support drug development.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This work demonstrated how advanced NLP can be leveraged to expedite the extraction and analyses of information for a few case examples in clinical pharmacology: (1) What is the trend for dose optimization in oncology? (2) What are the common covariates on pharmacokinetic (PK) parameters in population PK analyses for oncology drugs? (3) How frequent are PBPK analyses used to support regulatory submissions or labeling recommendations?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Advanced NLP is a powerful tool which can be leveraged to further support drug development by summarizing trends that could aid in performing risk–benefit assessments, informing clinical trial designs, monitoring pharmacovigilance and post-marketing surveillance, and facilitating steps toward personalized medicine.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Integrating advanced NLP into the clinical pharmacology workflow has the potential to increase efficiency in extracting impactful information that could ultimately optimize drug development.

The field of natural language processing (NLP) started in the 1940s, combining computer science with artificial intelligence (AI) to extract pertinent information efficiently and accurately from large amounts of natural language data. NLP has become increasingly ubiquitous and is embedded in various aspects of our daily lives, ranging from spell/grammar checks, online translators, search engines, and chatbots.[1] Despite the surge in usage in recent years, NLP has not been widely evaluated and utilized to support drug development.[2–4]

[1]Clinical Pharmacology, Genentech, Inc., South San Francisco, California, USA; [2]Computational Sciences, Genentech, Inc., South San Francisco, California, USA; [3]A2-Ai, Ann Arbor, Michigan, USA; [4]Capgemini America, Inc., New York, New York, USA. *Correspondence: Joy C. Hsu (hsu.joy@gene.com)

† Co-first authors.
‡ Co-senior authors.

Received September 26, 2023; accepted December 20, 2023. doi:10.1002/cpt.3161

The US Food and Drug Administration (FDA) has made a vast amount of documents from drug approval packages available to the public.[5] Several publications have manually mined these publicly available data. For example, Shi *et al*.[6] extracted evidence from drug labels to provide a resource for physicians on how to use drug products for treatment. Lu *et al*.[7] reviewed the Clinical Pharmacology and Biopharmaceutics Summary Basis of Approval (SBA) documents and drug labels for oncology new molecular entities (NMEs) to summarize learnings about approved doses in the drug label and their relationship with postmarketing requirements or commitments (PMRs/PMCs).

The SBA documents and drug labels contain valuable information that can be leveraged to support molecules under development and future strategic direction in drug development. However, manually summarizing these large documents can be time-consuming, which may deter scientists from using them for evidence syntheses. Even if manual curation is conducted, the process can still be tedious, time-consuming, and error-prone. Fortunately, NLP offers an opportunity to analyze unstructured text and enable efficient and accurate extraction of key information. By leveraging machine/deep learning technologies, advanced NLP can be trained to (1) retrieve relevant sentences with evidence using text classification (TC),[8] (2) recognize scientific evidence in key phrases or keywords by named entity recognition (NER), and question answering (QA),[9] and (3) extract entity relationships, such as the identification of adverse drug events by relationship extraction (RE).[10] With advanced training, NLP can efficiently and accurately survey large amounts of data, refine the most important items, and structure the data to allow for interpretation and further analyses.

Advanced NLP can be leveraged to expedite the extraction and analysis of information to help address an array of clinical pharmacology questions, inform clinical trial designs, and support drug development. To demonstrate how advanced NLP can be leveraged to support drug development, the following three clinical pharmacology use cases were selected as there was information previously curated manually,[7,11,12] which was ideal for NLP model development and assessment:

1. Dose optimization strategy in oncology: The dose optimization strategy was assessed for 112 NME applications for oncology drugs approved by the FDA from 2010 through the third quarter (Q3) of 2020. Data sources included the clinical pharmacology and biopharmaceutics SBA review documents, US package inserts (USPIs), and approval letters. The objectives were to summarize the dose selection rationale for registrational trials, the relationship between maximum tolerated dose (MTD) and the final label dose, the exposure-response (E-R) analyses conducted to support the optimal dose, and whether dose optimization-related PMRs/PMCs for NME applications are common for oncology drugs.
2. Common covariates on pharmacokinetic (PK) parameters in oncology: Significant covariates identified by the population PK (PopPK) analysis were assessed for NME oncology drugs approved from 2010 through Q3 of 2020 by the FDA. The effect of covariates on PK parameters summarized in SBA review documents and USPI were evaluated, and the most commonly used covariates were identified for both large and small molecules.
3. Physiologically-based PK (PBPK) analyses for regulatory review and product label: The impact and use of PBPK analysis to support NME applications were surveyed for NME that were approved by the FDA from 2013 through Q3 of 2020. Among the 328 NME across therapeutic areas, 72 approved drugs were identified to contain PBPK analysis results in the drug approval package, and their SBA review documents and USPI were evaluated.

METHODS

In-house NLP tool: Highlighting Annotation Web Kit

The NLP Text Highlighting and Annotation Web Kit (HAWK) is a customized AI tool that uses advanced deep learning NLP models and custom document-specific processing pipelines to retrieve and annotate documents more efficiently. In addition, HAWK also allows users to: (1) manually annotate evidence in a user-friendly interface, which enables quick generation of training data for NLP models and (2) implement final review of NLP result and approval by a domain expert, which is a crucial step to ensure the quality of the data for NLP model building as well as injection into downstream informatics systems.

Data preparation

To prepare the data for the NLP workflow (**Figure 1**), drug candidates were manually curated from the FDA website on NME drug and new biologic approvals.[13] Using New Drug Application numbers for those drug candidates, PDFs of the original SBA review documents, drug labels, and approval letters were systematically downloaded from the FDA website and fed into HAWK for data preprocessing and extraction.

In this workflow, SBA review documents, drug labels, and approval letters for 149 drug candidates approved between 2010 and Q3 of 2020 were prepared for the use cases described in this article. Among 149 drugs, 112 oncology drugs were assessed for their dose optimization strategy and common covariates on PK parameters. For the PBPK analyses use case, a total of 72 drugs (23 non-oncology and 49 oncology drugs) were analyzed.

Data preprocessing

Data conversion. Document processing starts with conversion of files from PDF to HTML format using Abbyy Finereader, which uses optical character recognition on the input PDF file to infer an HTML structure.[14] This step yields files where the text, tables, and figures from the original document are organized into an HTML tree, complete with properties for each node, such as an inferred font size, font weight, indentation, etc. This HTML file is the entry point for all document processing because contents of HTML files are organized and easily accessible.

Section extraction. To begin the process of information retrieval, documents are automatically pruned to only include text that is likely to contain key information. For example, when extracting information on exposure-efficacy (EE) relationships, the document is parsed to identify sections of text that are likely to contain EE-related information. All other sections are not processed. This step is referred to as "section extraction." The motivation behind this method is that documents can be large, and preselecting sections of text for processing can improve performance and speed.

Scientific documents are often organized into labeled sections, so an intuitive approach to identify relevant information would be to split a document into sections and select sections that have related labels. This is the foundation of our section extraction. We devised a set of rules to parse the HTML tree and identify section headers.

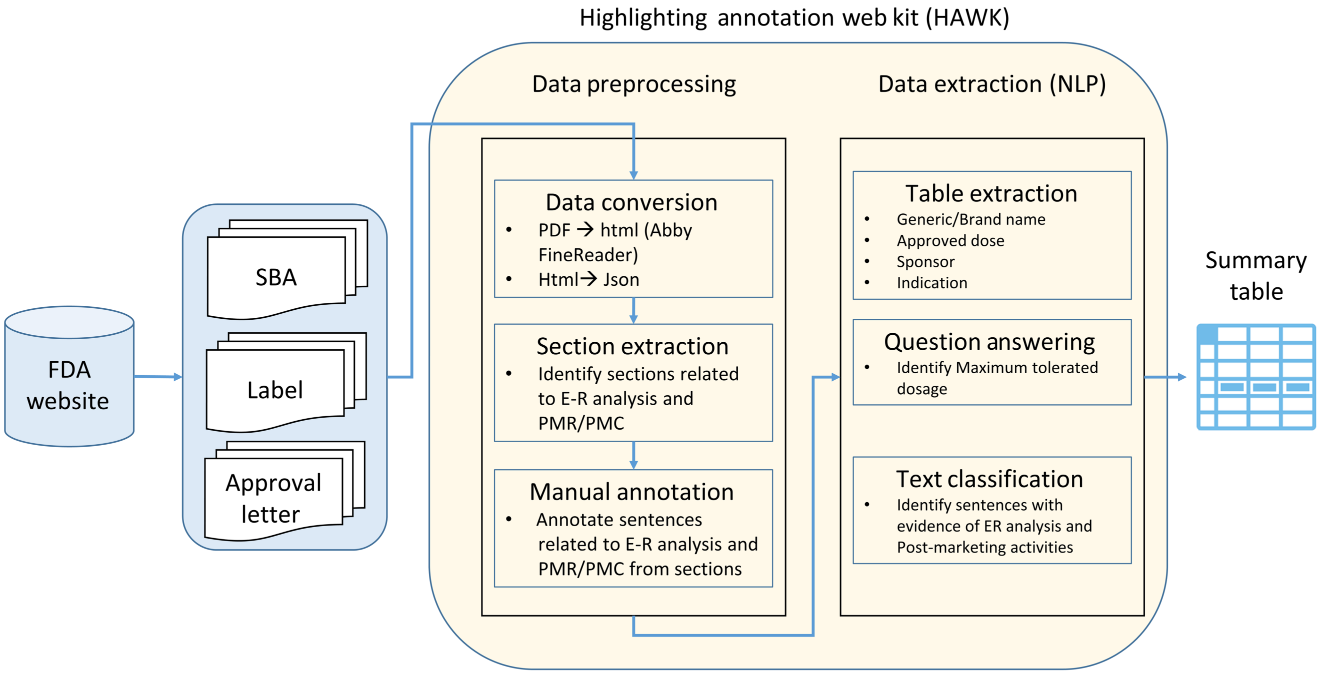

Figure 1 Overview of NLP workflow. The NLP workflow includes three components: (1) data preparation for collecting PDF files from the FDA website, (2) data preprocessing for NLP model building, and (3) automated data extraction from tables and text for final review. E-R, exposure-response; FDA, US Food and Drug Administration; NLP, Natural language processing; PMR/PMC, postmarketing requirements or commitments; SBA, summary basis of approval.

Data extraction *Question answering*. Some metadata fields, such as MTD, are less likely to be reported in a consistent fashion and are often contained within the text of the document itself. In these cases, the pattern matching approach is inadequate and an NLP approach is needed.[15] QA is a common NLP task in deep learning, which involves providing a question to a trained deep learning model, as well as providing a body of text (also referred to as "context") that potentially contains the answer. In this task, a pre-trained BioBERT QA model from Huggingface was used.[16] The model processes the context and extracts the span of text that it calculates to most likely contain the answer to the question. For example, to extract information on MTD, the QA model will identify all paragraphs containing MTD-related information. Subsequently, the extracted paragraphs are used as the context for the question: "What is the maximum tolerated dose or MTD related to dose selection?"

Text classification. Machine learning methods were used to automatically annotate sentences for categories of interest, and labeled data were required to train the model to annotate. The categories of interest, which are related to dose selection, include EE analysis, exposure-safety (ES) analysis, PMR, and PMC.

- Acquiring labeled data: Training data were collected by manually annotating key sentences for categories of interest from a collection of documents. The output of this step is a CSV file for each category, with columns containing each sentence from the document that is extracted for that particular category, the document name, and the sentence label. Sentences are labeled 1 if they are annotated and 0 if they are not. For each category, 41 out of 112 documents were manually annotated, resulting in several thousand sentence samples and labels per category.
- Training TC models: The labeled sentences were then used to train machine learning models to perform automatic annotation in a way that mimics how humans would annotate a document. The specific model architecture used was a variant of the popular BERT architecture called ERNIE-BERT.[17] In our case, a separate model was trained for each category.
- Predicting annotations on new documents: Once a model is trained for each category, they could be used to predict annotations of new documents to enable automatic annotation. Given a batch of unannotated documents, a python script could be executed to extract sections for each category, split the section into individual sentences, and subsequently predict whether the sentence should be annotated. The output of this step is the same as the output of a manually annotated document, which is a .json file containing all annotated sentences.

Details of the NLP workflow are described in Methods S1.

RESULTS

Dose optimization strategy in oncology

Lu *et al.*[7] conducted a comprehensive survey of the label dose for NME applications for oncology drugs approved by the FDA from 2010 through Q1 of 2015. Using clinical pharmacology and biopharmaceutics SBA review documents, USPI, and approval letters, a total of 41 oncology drugs were manually curated. The dose optimization use case highlights the advantage of leveraging an advanced NLP platform to extract and summarize the dose selection rationale for registrational trials, the relationship of the MTD and the final label dose, the PMR or PMC related to dose optimization activities, and typical E-R analysis conducted for oncology drugs approved by the FDA from 2010 through Q3 of 2020.

To enable NLP for automatically extracting the evidence for EE analysis, ES analysis, and PMR/PMC related to dose selection, three NLP models (model 1 for EE, model 2 for ES, and model 3 for PMR/PMC) were used. The 41 drugs surveyed by Lu *et al*.[7] were used to train the models as the first step. Among the 41 drugs, we identified 29 positive sentences for EE, 147 positive sentences for ES, and 68 positive sentences for PMR/PMC. Subsequently, three TC models were developed to enable automated extraction of sentences containing evidence of E-R analyses for safety and efficacy and postmarketing requirements related to dose optimization activities for NME applications for oncology drugs approved by the FDA from the second quarter (Q2) of 2015 through Q3 of 2020. In addition, QA was applied to the paragraphs containing MTD evidence, and sentences containing evidence on the relationship of the MTD and the final label dose were identified.

Summary. A total of 112 NME applications for oncology drugs were approved by the FDA from 2010 through Q3 of 2020, 71 of which were approved after Q1 of 2015. Drug development in oncology using MTD was common; 33% proposed the MTD as the labeled drug dose, of which 86% were small molecules (**Table 1**). Of those oncology drugs that proposed the MTD as the labeled drug dose, 27% (10/37) had a PMR or PMC issued for dose optimization activities. For those drugs where the MTD was not identified or was not proposed as the labeled drug dose, 84% (63/75) conducted E-R analyses to support the dose rationale. Approximately 15% of the oncology drugs approved between 2010 and Q3 of 2020 were issued PMR or PMC to identify any alternative dose which may confer a better risk–benefit profile.

The number of small and large molecule oncology drugs that were approved from Q2 of 2015 to Q3 of 2020 almost doubled compared with those approved between 2010 and Q1 of 2015 (**Table 1**). In addition, the number of oncology drugs that were granted accelerated approval (AA), breakthrough therapy (BT), fast track (FT), or orphan drug (OD) designation between Q2 of 2015 to Q3 of 2020 more than doubled compared with those approved between 2010 and Q1 of 2015 in each category (**Table 2**). Of the oncology drugs that were granted AA, BT, FT, or OD designation from 2010 through Q3 of 2020, the percentage of approved drugs that were issued PMR or PMC related to dose justification or optimization decreased under each designation category after Q1 of 2015 compared with that from 2010 through Q1 of 2015 (**Table 2**).

Compared with results presented in Lu *et al*.,[7] the NLP approach was able to come to the same conclusion regarding the trends in the MTD dosing approach and PMRs/PMCs related to dose optimization for oncology NMEs approved from 2010 through Q3 of 2020. Detailed results for the dose optimization use case are summarized in **Table S1**.

Common covariates on PK parameters in oncology

To our knowledge, a comparison of PopPK covariates for small and large oncology molecules has not been summarized. This use case builds on a previous study by Bajaj *et al*.[11] which

Table 1 Statistics and categorization of approved oncology drugs from 2010 through Q3 of 2020

	Small molecule					Large molecule				
	TKIs (*n*)	Non-KI small molecule targeted drugs (*n*)	Others (radioactive therapeutic agent and chemotherapy) (*n*)	Total (*n*)	MTD proposed as label dose (%) [# of drug with MTD proposed as label dose/total # of small molecules approved during each time frame]	Antibody drug conjugates (*n*)	Monoclonal antibodies (*n*)	Others (enzyme and fusion protein) (*n*)	Total (*n*)	MTD proposed as label dose (%) [# of drug with MTD proposed as label dose/total # of large molecules approved during each time frame]
Q1 2010 – Q1 2015	16	8	4	28	64% [18/28]	2	9	1	13	23% [3/13]
Q2 2015 – Q3 2020	30	17	4	51	27% [14/51]	9	10	1	20	10% [2/20]
Q1 2010 – Q3 2020	46	25	8	79	40% [32/79]	11	19	2	33	15% [5/33]

MTD, maximum tolerated dose; Non-KI, non-kinase inhibitor; Q1/2/3, quarter 1/2/3; TKI, tyrosine kinase inhibitors.

Table 2 Summary of drugs that have a PMR or PMC related to dose optimization across different designations

	Types of approval			
	AAs (# of drug with AA that has a PMR related to optimal dose/# of AA drugs)	BT (# of drug with BT that has a PMR related to optimal dose/# of BT drugs)	FT (# of drug with FT that has a PMR related to optimal dose/# FT of drugs)	OD (# of drug with OD that has a PMR related to optimal dose/# of OD drugs)
Q1 2010 – Q1 2015	2/11 (18%)	4/16 (25%)	1/11 (9%)	3/18 (17%)
Q2 2015 – Q3 2020	3/30 (10%)	4/39 (10%)	1/27 (4%)	4/54 (7.0%)
Q1 2010 – Q3 2020	5/41 (12%)	8/55 (15%)	2/38 (5%)	7/72 (10%)

AA, accelerated approvals; BT, breakthrough therapy; FT, fast track; OD, orphan drug; PMC, post marketing commitments; PMR, post marketing requirements; Q1/2/3, quarter 1/2/3.

manually curated PopPK covariates for 25 monoclonal antibodies in the oncology space. Here, we present findings from 82 NME applications for oncology drugs (53 small and 29 large molecules) which were approved by the FDA between 2010 and Q3 of 2020. In this use case, NER and RE were utilized to identify which covariates affect different PK parameters. In addition to the relationship between PK parameters and covariates, the statistical and clinical significance of these relationships were extracted and assessed.

Summary. Using our NLP technologies (NER and TC) for mining SBA documents via HAWK, the covariates were extracted for the 82 oncology NMEs that have PopPK information. The three most common covariates identified for both large and small molecules are body weight, albumin, and sex (**Figure 2**). Approximately 20% of the large molecules evaluated mention the impact of body weight on exposure (**Figure 3**). Creatinine clearance, age, body surface area (BSA), and race are common covariates for small molecule drugs but not as common for large molecule drugs (**Figure 3**). Clearance appears to be the PK parameter that is most affected by the identified covariates, followed by volume (**Figure 3**).

PBPK analyses for regulatory review and product label

With increasing use of PBPK modeling in drug development, the impact of PBPK on regulatory reviews and drug labels was previously evaluated by Yoshida *et al.*[12] for NMEs approved from January 2013 through August 2016. The current use case extends this work to survey the inclusion of PBPK modeling in the clinical pharmacology and biopharmaceutics SBA review documents for drugs approved between 2013 and Q3 of 2020. Using the NLP approach, a total of 72 drugs which included PBPK-related keywords were identified from Q1 of 2013 to Q3 of 2020. The keywords utilized include terms and synonyms related to PBPK applications (e.g., Simcyp, GastroPlus, Physiolab, etc.), PK parameters (e.g., area under the curve (AUC), clearance (CL), maximum concentration (C_{max}), etc.), and Enzyme/Transporter names (e.g., CYP3A4, UGT1A1, etc.).

Among the 72 drugs, 27 were utilized as the training dataset to train the TC models (17 between January 2013 and August 2016 and 10 between 2017 and Q3 2020).[12] HAWK was used for manual annotation to identify different categories of information on "Detail of PBPK Predicted Scenarios," "PBPK Simulation Result," "Impact/outcome on drug label," and "Dataset or strategy for model development/validation," which lead to the identification of 150, 233, 147, and 163 positive sentences, respectively, for these four TC models. Automatic extraction using the 4 developed TC models was then performed for the remaining 45 drugs.

Summary. As shown in **Figure 4**, the leading use for PBPK models is for predicting the effect of metabolic enzyme and/or transporter mediated drug–drug interactions (DDI), followed by predicting absorption and evaluating the effects of organ impairment. The number of PBPK applications increases over the years across all areas (**Figure 4**). This increase could be attributed to the ongoing efforts in the field for PBPK model verification in comparison to those prior to 2016.[18–20]

DISCUSSION

Dose optimization strategy in oncology

Historically, the dosing strategy for oncology drugs relied heavily on the identification of the MTD to deliver safe and efficacious treatments quickly to patients with cancer who oftentimes have limited treatment options. It was believed that dose and therapeutic effect are linked. Therefore, the MTD would provide the greatest therapeutic benefit, and using MTD as the labeled dose would provide the fastest development path. However, dose optimization remains a major challenge today for the development of oncologic therapeutics due to disease heterogeneity, additive toxicity from combination therapy, and drug response variability, among many other factors.

Between 2010 and Q1 of 2015, ~50% (21/41) of the approved oncology drugs had proposed label dose at MTD, of which 43% (9/21) had a PMR or PMC related to dose justification or optimization, including clinical trials to study alternative doses to the approved label dose, or to conduct additional E-R analyses to justify the label dose.[7] These PMRs or PMCs suggested that the MTD approach was inadequate to find the optimal dose. The more recent trend appears to move away from developing drugs at MTD, as only 23% (16/71) of oncology drugs approved between Q2 of 2015 and Q3 of 2020 proposed the MTD as the labeled drug dose. Not only was there a decrease in the number of drugs approved at MTD as labeled dose, the percentage of drugs that were issued a PMR or PMC related to dose justification or optimization also decreased (2010-Q1 of 2015: 27% (11/41), Q2 of 2015-Q3 of 2020: 8% (6/71)). Moreover, although the number of oncology drugs that were granted AA, BT, FT, or OD designation more

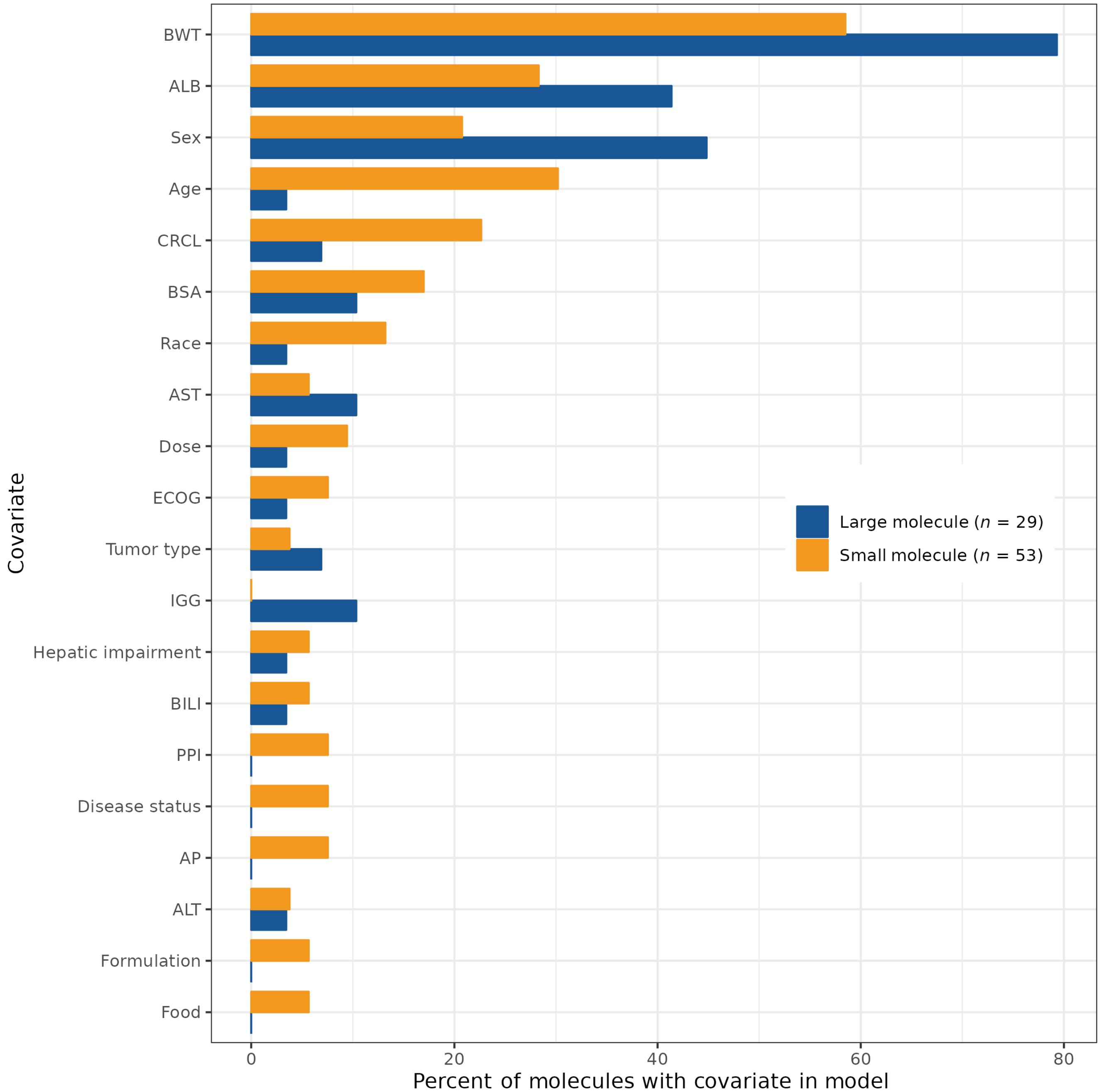

Figure 2 Percent of molecules with covariates included on PK parameters to explain variability using PopPK analyses for oncology drugs approved by the FDA between 2010 and Q3 of 2020, by molecular size. ALB, albumin; ALT, alanine transaminase; AP, alkaline phosphatase; AST, aspartate aminotransferase; BILI, bilirubin; BSA, body surface area; BWT, body weight; CRCL, creatinine clearance; ECOG, Eastern Cooperative Oncology Group performance status; FDA, US Food and Drug Administration; IGG, immunoglobulin G; PK, pharmacokinetic; PopPK, population pharmacokinetic; PPI, proton pump inhibitor.

than doubled between Q2 of 2015 and Q3 of 2020, there was a decreasing trend in the PMRs or PMCs issued in each category that were related to dose optimization. This may be attributed to the increased utilization and integration of the model-informed drug development (MIDD) approach.[21]

With the launch of Project Optimus, an initiative to reform and improve dose optimization and selection in oncology drug development, by the FDA Oncology Center of Excellence in 2021, there are additional opportunities to move toward a dose optimization paradigm that emphasizes selection of a dose or doses that maximizes the efficacy, safety, and tolerability of a drug.

Common covariates on PK parameters in oncology

NLP allows for systematic and efficient identification of covariates that are repeatedly identified as being significant across large and small molecule drugs. Our findings are consistent with previously published work by Bajaj *et al.*[11] which also identified body weight, sex, and albumin as significant covariates on CL, body weight, and

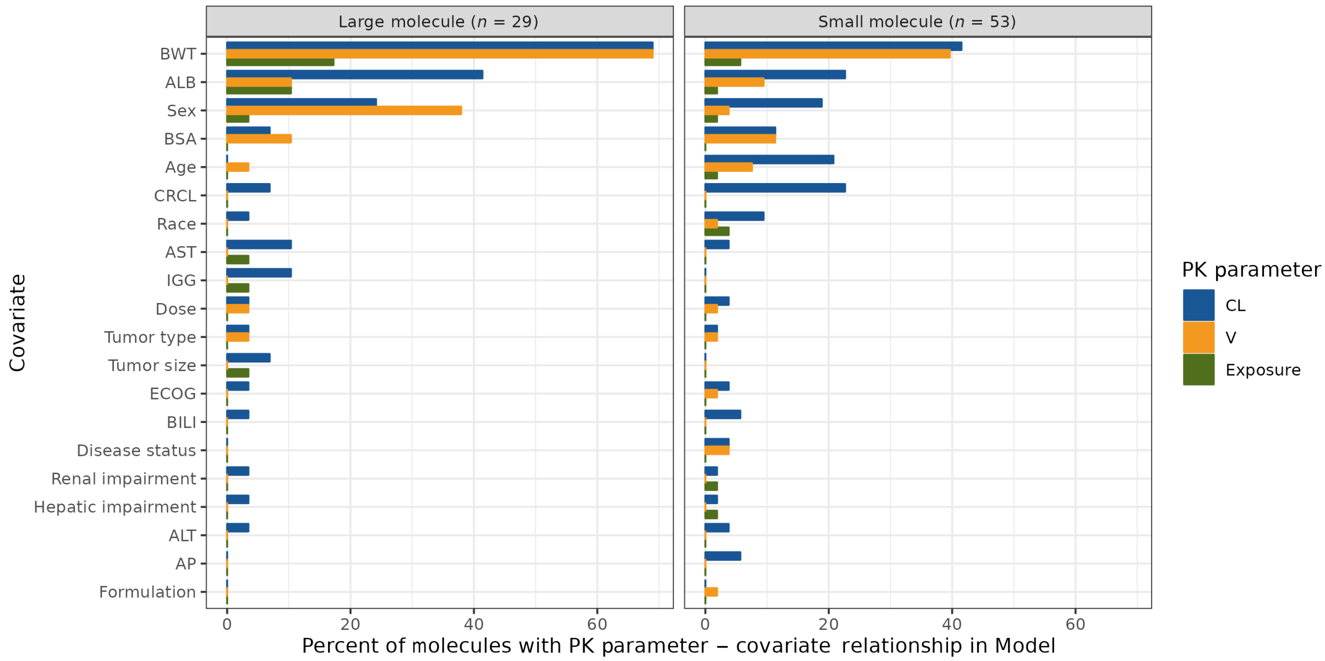

Figure 3 Percent of molecules with covariates included on PK parameters to explain variability using PopPK analyses for oncology drugs approved by the FDA between 2010 and Q3 of 2020, by PK parameter and by molecular size. ALB, albumin; ALT, alanine transaminase; AP, alkaline phosphatase; AST, aspartate aminotransferase; BILI, bilirubin; BSA, body surface area; BWT, body weight; CRCL, creatinine clearance; CL, clearance; ECOG, Eastern Cooperative Oncology Group performance status; FDA, US Food and Drug Administration; IGG, immunoglobulin G; PK, pharmacokinetic; PopPK, population pharmacokinetic; V, volume.

sex as significant covariates on volume (**Figure 3**). Comparing the covariates for large and small molecules, there is an overlap for certain covariates, such as body weight, albumin, and sex. However, there is also a large number of covariates consistently identified for small molecules (compared with large molecules) such as creatinine CL, age, food effect, use of proton pump inhibitors, and race, all of which are factors and variables considered and accounted for during development of small molecule drugs (**Figure 2**). Finally, as expected, these covariates are found to have an impact on CL, in line with how age, kidney function, BSA, and race may affect drug elimination (**Figure 3**).

The use case presented here summarizes significant covariates across 82 small and large molecule oncology drugs (53 small and 29 large molecules) approved between 2010 and Q3 of 2020. These findings have great utility in informing the design of future clinical trials as well as help make strides toward personalized medicine. With respect to the former, this work can help inform what type of data should be collected in future clinical studies, how patients may need to be stratified, and/or which inclusion/exclusion criteria should be implemented to ensure a wide range of patients are enrolled to allow assessment of a covariate of interest (i.e., obese vs. non-obese patients because weight is consistently a significant

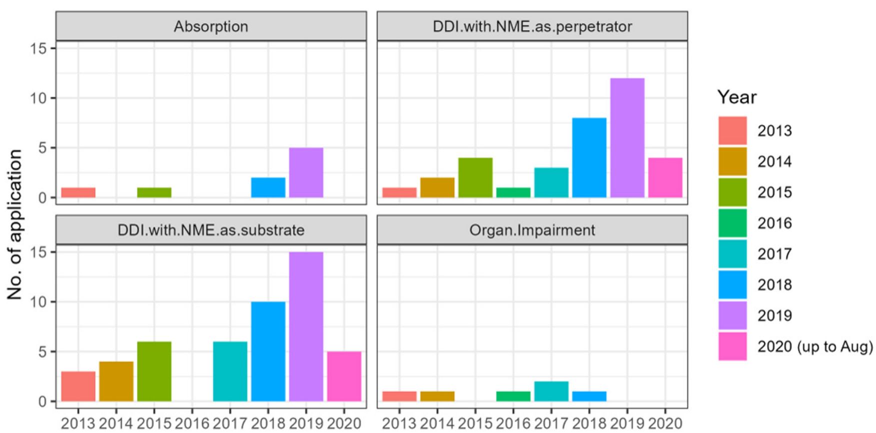

Figure 4 Areas of physiologically-based pharmacokinetic (PBPK) application for drugs approved by the US Food and Drug Administration between 2013 and Q3 of 2020. DDI, drug–drug interaction; w/, with; NME, new molecular entities.

covariate across large and small molecules). Additionally, using NLP to systematically identify which covariates consistently have a significant and/or clinical impact on drug absorption, distribution, exposure, and/or elimination across various large and small molecules help us make strides toward personalized medicine.

PBPK analyses for regulatory review and product labels

PBPK models have increasingly been used in regulatory submissions in recent years, especially for labeling recommendations.[22–24] Consistent with the previous evaluation by Yoshida *et al*.,[12] the regulatory impacts of PBPK analysis mainly focus on the assessment of potential metabolism-based DDI and absorption-related PK changes for NMEs that were approved between August 2016 and Q3 of 2020. It is worth noting that there is an increase in the number of NMEs for which PBPK analyses were included in the submissions, especially from 2016 through 2020. This may be attributed to the increased utilization and integration of the MIDD approaches in recent years. In addition, the accumulated experiences in PBPK analyses over the years has likely helped to establish the level of confidence that is required to inform regulatory decisions.[25]

Challenges and future opportunities

While we have demonstrated several advanced NLP use cases to address clinical pharmacology questions efficiently, there are several challenges. First and foremost, postprocessing steps are required to organize information extracted by NLP methods to produce interpretable and meaningful results. Without any postprocessing, the raw output from NLP lacks sufficient structure for direct analysis and interpretation. Second, it is essential to provide specific terminologies in order to extract data that is tailored for each search. Otherwise, the results extracted by NLP will likely be too broad and may not be useful or meaningful. Third, when the topic of interest includes multiple terminologies, it may be challenging to extract complete information if not all synonyms are provided. For example, "exposure-response" can be written as "exposure-efficacy," "exposure-safety," "exposure-efficacy/safety," or "ER," to name a few. For terms that lack standardization, it is important to include alternative phrases or wording to ensure extraction of pertinent information. Last, text included in figures and tables cannot be mined due to differences in formatting for each figure and table. This is a challenging task for current NLP technologies, as valuable information could reside in the figures and tables. Potentially, this can be resolved by using a novel Large Language Model (LLM) built specifically for figures and tables. Although LLM can provide an efficient way for information retrieval from unstructured documents, it is also not fully matured yet. It still relies highly on asking specific questions and providing well-crafted prompts.

As demonstrated in the use cases, advanced NLP can be leveraged to expedite the extraction and analyses of information from regulatory review documents and product labels to help address important clinical pharmacology questions. Once an NLP model is developed, it can be used to address various questions. Going forward, close collaboration between Clinical Pharmacologists and Computational Scientists are needed to apply NLP effectively into the Clinical Pharmacology workflow to help advance drug development. Expanding the NLP platform, such as including PubMed abstracts/papers, journals and conference abstracts included in Embase and the University of Washington's Drug Interaction Database, will further enhance its ability to mine valuable information. There are other FDA resource-derived databases, such as Side Effect Resource (SIDER)/SIDER2 for adverse events that could also be integrated into clinical pharmacology analysis if warranted. Although it has not yet been adopted widely, advanced NLP is a very powerful tool to further support drug development by summarizing trends that could aid in performing risk–benefit assessment, informing clinical trial designs, monitoring pharmacovigilance and postmarketing surveillance, and facilitating steps toward personalized medicine. Hence, integrating advanced NLP into the clinical pharmacology field has the potential to increase efficiency in extracting impactful information that could ultimately optimize drug development and lower costs to patients.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors thank Drs Amita Joshi and Hongmei Huang for their helpful discussions and feedback.

FUNDING

This work is supported by Genentech Inc.

CONFLICT OF INTEREST

J.C.H., M.W., C.K., B.V., Y.T.(Kayla)L., A.J., K.Y., J.Y.J., C.L., B.C., and B.W. are employees of Genentech Inc. and hold stocks/shares in F. Hoffmann-La Roche Ltd. S.K. is a previous employee of Genentech Inc. J.G. is an employee of Capgemini America, Inc.

AUTHOR CONTRIBUTIONS

J.C.H., M.W., C.K., B.V., K.L., A.J., and B.W. wrote the manuscript. J.C.H., M.W., C.K., J.Y.J., C.L., B.C., and B.W. designed the research. J.C.H., M.W., C.K., B.V., K.L., A.J., K.Y., S.K., J.G., and B.W. performed the research. J.C.H., M.W., C.K., B.V., K.L., A.J., K.Y., S.K., and B.W. analyzed the data.

1. Hagiwara, M. *Real-World Natural Language Processing: Practical Applications with Deep Learning* (Simon & Schuster: New York, 2021).
2. Bitterman, D.S., Miller, T.A., Mak, R.H. & Savova, G.K. Clinical natural language processing for radiation oncology: a review and practical primer. *Int. J. Radiat. Oncol. Biol. Phys.* **110**, 641–655 (2021).
3. Tulipano, P.K. *et al*. Natural language processing in the molecular imaging domain. *AMIA Annu. Symp. Proc.* **2005**, 1143 (2005).
4. Masanz, J. *et al*. Open source clinical NLP – more than any single system. *AMIA Jt Summits Transl Sci Proc* **2014**, 76–82 (2014).
5. Ladanie, A., Ewald, H., Kasenda, B. & Hemkens, L.G. How to use FDA drug approval documents for evidence syntheses. *BMJ* **362**, k2815 (2018).

6. Shi, Y., Ren, P., Zhang, Y., Gong, X., Hu, M. & Liang, H. Information extraction from FDA drug labeling to enhance product-specific guidance assessment using natural language processing. *Front Res Metr Anal* **6**, 670006 (2021).
7. Lu, D. *et al*. A survey of new oncology drug approvals in the USA from 2010 to 2015: a focus on optimal dose and related postmarketing activities. *Cancer Chemother. Pharmacol.* **77**, 459–476 (2016).
8. Kolchinsky, A., Lourenco, A., Wu, H.Y., Li, L. & Rocha, L.M. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS One* **10**, e0122199 (2015).
9. Lee, J. *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
10. Wu, H.Y. *et al*. Translational knowledge discovery between drug interactions and Pharmacogenetics. *Clin. Pharmacol. Ther.* **107**, 886–902 (2020).
11. Bajaj, G., Suryawanshi, S., Roy, A. & Gupta, M. Evaluation of covariate effects on pharmacokinetics of monoclonal antibodies in oncology. *Br. J. Clin. Pharmacol.* **85**, 2045–2058 (2019).
12. Yoshida, K., Budha, N. & Jin, J.Y. Impact of physiologically based pharmacokinetic models on regulatory reviews and product labels: frequent utilization in the field of oncology. *Clin. Pharmacol. Ther.* **101**, 597–602 (2017).
13. FDA. Compilation of CDER New Molecular Entity (NME) Drug and New Biologic Approvals <https://www.fda.gov/drugs/drug-approvals-and-databases/compilation-cder-new-molecular-entity-nme-drug-and-new-biologic-approvals> [cited 2023].
14. Tafti, A.P. *et al*. *OCR as a service: an experimental evaluation of Google docs OCR, tesseract, ABBYY FineReader, and Transym*. 735–746 (Springer International Publishing, Cham, 2016).
15. Lopez, M.M. & Kalita, J. Deep Learning applied to NLP (2017) https://doi.org/10.48550/arXiv.1703.03091. Cited September 20, 2023.
16. ktrapeznikov biobert_v1.1_pubmed_squad_v2 model <https://metatext.io/models/ktrapeznikov-biobert_v1.1_pubmed_squad_v2>. Accessed June 2, 2023.
17. Zhang Z., et al. ERNIE: Enhanced Language Representation with Informative Entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy: Association for Computational Linguistics; 2019. p. 1441–1451.
18. Kuemmel, C. *et al*. Consideration of a credibility assessment framework in model-informed drug development: potential application to physiologically-based pharmacokinetic modeling and simulation. *CPT Pharmacometrics Syst. Pharmacol.* **9**, 21–28 (2020).
19. Peters, S.A. & Dolgos, H. Requirements to establishing confidence in physiologically based pharmacokinetic (PBPK) models and overcoming some of the challenges to meeting them. *Clin. Pharmacokinet.* **58**, 1355–1371 (2019).
20. Shebley, M. *et al*. Physiologically based pharmacokinetic model qualification and reporting procedures for regulatory submissions: a consortium perspective. *Clin. Pharmacol. Ther.* **104**, 88–110 (2018).
21. Mitra, A. & Wang, Y. Applications of model informed drug development (MIDD) in drug development lifecycle and regulatory review. *Pharm. Res.* **39**, 1663–1667 (2022).
22. Jean, D. *et al*. Development of best practices in physiologically based pharmacokinetic modeling to support clinical pharmacology regulatory decision-making-a workshop summary. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 1271–1275 (2021).
23. Zhang, X. *et al*. Application of PBPK modeling and simulation for regulatory decision making and its impact on US prescribing information: an update on the 2018-2019 submissions to the US FDA's Office of Clinical Pharmacology. *J. Clin. Pharmacol.* **60**(Suppl 1), S160–S178 (2020).
24. Lin, W., Chen, Y., Unadkat, J.D., Zhang, X., Wu, D. & Heimbach, T. Applications, challenges, and outlook for PBPK modeling and simulation: a regulatory, industrial and academic perspective. *Pharm. Res.* **39**, 1701–1731 (2022).
25. Bischoff, K.B., Dedrick, R.L., Zaharko, D.S. & Longstreth, J.A. Methotrexate pharmacokinetics. *J. Pharm. Sci.* **60**, 1128–1133 (1971).