Chem-Bio Informatics Journal, Vol.23, pp.1-6(2023)

Difficulties and prospects of data curation for ADME *in silico* modeling

Tsuyoshi Esaki[1*], Kazuyoshi Ikeda[2,3]

[1] *Data Science and AI Innovation Research Promotion Center, Shiga University, 1-1-1 Banba, Hikone, Shiga, 522-8522, Japan*
[2] *Keio University Faculty of Pharmacy, 1-5-30 Shibakoen, Minato-ku, Tokyo 105-8512, Japan*
[3] *HPC-and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 4230-0045, Japan*

**E-mail: tsuyoshi-esaki@biwako.shiga-u.ac.jp*

(Received October 22, 2022 ; accepted December 23, 2022; published online January 13, 2023)

Abstract

The cost and time required for drug discovery must be reduced. Recent *in silico* models have focused on accelerating seed compound discovery based solely on chemical structure. Estimating pharmacokinetic characteristics, including absorption, distribution, metabolism, and excretion (ADME), is essential in the early stage of drug discovery. Therefore, *in silico* models have used artificial intelligence (AI) techniques to predict the ADME properties of potential compounds. Large experimental data are necessary when constructing *in silico* models for ADME prediction. However, it remains difficult for one pharmaceutical company or academic laboratory to collect enough data for modeling. Therefore, collecting data from open databases with the assistance of dry scientists is one of the most effective strategies utilized by researchers. However, incorrect values are occasionally included in open databases because of human errors. Furthermore, to construct high-performance ADME *in silico* models, data curation must include not only chemical structure but also experimental conditions, which requires expert knowledge of pharmacokinetic experiments. Trials to ease the difficulties of data curation have been developed as reported. These tools enable the effective collection and checking of published data. Additionally, they accelerate collaboration between dry and wet scientists, enabling them to collect vast amounts of data to construct high-performance and widespread chemical space ADME *in silico* models. Collecting much accurate data for constructing ADME *in silico* models is an expectation of the new era of efficient drug discovery when entirely using AI technology.

Key Words: ADME, data curation, *in silico* modeling, prediction

Area of Interest: *In silico* drug discovery

1. Introduction

The cost and time required for drug discovery must be reduced. Therefore, *in silico* models, which estimate features of chemical compounds using informatic approaches, have recently focused on efficiently accelerating finding seed compounds solely using chemical structure. Estimating pharmacokinetic characteristics, including absorption, distribution, metabolism, and excretion (ADME), are essential because ADME affects the estimate of a compound's drug-likeness and the calculation of administered dosage. Poor pharmacokinetics of compounds can cause failures in the late stages of drug discovery[1]. For these reasons, *in silico* models using artificial intelligence (AI) techniques have been expected to predict *in vitro* ADME properties of lead compounds in the early stage of drug discovery.

A huge amount of experimental data is required to construct *in silico* models to predict ADME properties. High throughput screening (HTS) facilitates the accessibility of data and efficiency in obtaining many *in vitro* measurements efficiently. However, it remains difficult for one pharmaceutical company or academic laboratory to collect enough experimental data for high-performance *in silico* modeling. Prediction models constructed using a small amount of data result in low accuracy and restricted applicability in chemical space models.

Therefore, one of the most effective strategies when constructing *in silico* models is collecting large amounts of experimental data from open databases, such as ChEMBL [2], a manually curated database with drug-like properties, with the assistance of the dry scientists. However, there are several hurdles to collecting many measurements using an open database because they occasionally include incorrect information [3]. Furthermore, it is difficult to obtain assurance when collecting experimental data measured by single protocols. These miscellaneous data result in a reduced quality of the dataset and decrease the performance of constructed models [4, 5].

To publish high-quality data for model development, "data curation" is necessary. However, curation has typically been performed manually. Thus, it is expensive and time-consuming, and incorrect values are occasionally included. Therefore, to overcome these hurdles, the use of helpful curation tools is essential to reduce the need for manual procedures and to accelerate collaboration between dry and wet scientists. Collecting copious amounts of accurate data for constructing ADME *in silico* models is an expectation of the new era of efficient drug discovery when solely using AI technology.

2. Difficulty of data curation for public databases

Data curation, which involves carefully checking various types of information (such as compound structure, isomers, and experimental condition) stored for easy use by researchers, is necessary when publishing experimental data extracted from research articles or patents for use in cheminformatics. However, there are many obstacles to collecting massive amounts of accurate data for curation. Manual curation is costly and time-consuming. Furthermore, it is difficult to maintain up-to-date datasets by periodically checking the rapidly growing volume of publications [4]. To overcome these difficulties, in-house curators collect pharmaceutical data for entry into ChEMBL [2]. The curators review published data, and then extract and summarize them with a description of the experiment that can be referred to when gathering similar experimental data for data analysis. However, curating ADME data has two problems: (1) mistakes caused by human error and (2) lack of expert knowledge of pharmacokinetic experiments.

2.1 Mistakes caused by human error

People perform the manual curation of the data. Therefore, it is inevitable that data that differ from the original are overlooked. Information that includes molecular structure and activity values is likely to include several errors [3]. It is necessary to confirm whether the collected data have abnormal data, such as impossible structures, values, or inappropriate protocols. When collecting high-quality data, a first helpful step is to check for anomalous data in the original articles.

In the area of quantitative structure-activity relationship (QSAR), precise chemical structure is the most critical information. Therefore, curations for chemical data are focused on the check for chemical structures. Thus, several tools for chemical data curation have been developed [4]. A workflow constructed using KNIME [6] semi-automatically curates data, mainly chemical structure cleaning and data standardization. These procedures were reported for QSAR research [7]. To reduce the manual procedures in data curation and the associated human errors, it is practical to use these automation tools for data-driven drug discovery.

2.2 Lack of expert knowledge of pharmacokinetic experiments

The chemical structure and experimental conditions for data measurements are also essential factors when collecting ADME data for *in silico* prediction. ADME experiments have various *in vitro* pharmacokinetic properties, such as aqueous solubility, metabolic stability, membrane permeability, and plasma protein binding. As an ADME experiment example, metabolic stability is one of the most varied properties of experimental protocols. Metabolic stability, as measured using liver microsomes, varies substantially across companies and even between different laboratories within the same company [8]. The protocols to test the metabolic stability of liver microsomes have various conditions, such as compound concentration, incubation time, microsomal protein concentration, use of recombinant, use of nicotinamide adenine dinucleotide phosphate, and cut-off criteria.

When making high-accuracy and widespread chemical space models, researchers who construct ADME *in silico* models are required to select data measured in the target protocol and reject data obtained using different protocols. However, to accurately perform the selection and rejection processes, expert knowledge of pharmacokinetic experiments is required (Table 1). The amount of data that must be extracted as target protocols to form a proper dataset for ADME model construction is remarkable. These problems hinder the progress toward curating ADME data.

It is difficult to train dry scientists to conduct pharmacokinetic experiments. Therefore, it is efficient to consult wet scientists, such as medicinal chemists, regarding the essential points of *in vitro* ADME experiments. Accelerating collaboration between dry and wet scientists enables them to collect vast amounts of data to construct high-performance and widespread chemical space ADME *in silico* models.

Table 1. Example of requirements for checking conditions and units in protocols of a representative *in vitro* ADME experiment

Experiment	Condition	Unit
Aqueous solubility	Temperature, pH, usage of dimethylsulphoxide	M (mol/L), ug/mL, g/L, mg/L
Metabolic stability	Compound concentration, recombinant, species (e.g., *Homo sapiens*, *Rattus norvegicus*, *Macaca mulatta*), presence of cofactor, measured time, sample (microsome, hepatocyte, or plasma)	μL/min/mg, mL/min/mL, L/h
Membrane permeability	Cell type (Caco-2, LLC-PK1, MDCK1, PAMPA), the direction of penetration (from apical to basolateral or from basolateral to apical), compound concentration, incubation time, presence of cofactor	10^{-6} cm/s, μcm/s, 10^{-5} cm/s

3. Tools for data curation

Increasing the usable data is helpful for *in silico* model construction. However, it is difficult for individual researchers to manually curate a huge amount of published information. Therefore, trials to ease the difficulties of data curation have been developed as reported in previous articles. Kim and collaborators developed a KNIME tool to curate and prepare HTS data using KNIME [9]. To decrease the cost of curation, ChEMBL developed a method that distinguishes whether data are ChEMBL-like or not [10]. Supporting data-driven research to fully use published data has been necessary for the *in silico* approach. Thus, the development of tools that enable the effective collection and checking of published data is increasingly necessary. Similarly, it is also essential to use published data efficiently to construct ADME *in silico* models. Data curation, including experimental protocols, has become a necessary step. The development of curation tools for ADME experimental conditions is anticipated.

4. Conclusion

Estimating *in vitro* ADME properties of lead compounds early in the drug discovery stage is becoming increasingly critical. To realize this, it is necessary to use large quantities of data while maintaining high quality. Because of the usefulness of curated data, there has been an increase in the number of databases and the accumulation of curated datasets [11] [12] [13] [14] [15]. Curator's names are posted in a pharmacokinetics database, PK-PD, which implies that the importance of curation has increased. However, these datasets are not enough to predict ADME properties comprehensively and are inadequate to construct a local *in silico* model focused on a specific target. Therefore, curating collected data is an unavoidable step in the era of *in silico* drug discovery. For the new era, which accelerates drug discovery using the AI approach, reconstructing open data into usable high-quality datasets through data curation will be required.

Acknowledgments

This work was supported by JSPS KAKENHI (Grant Number: 20K16075). We thank Editage (www.editage.jp) for English language editing.

References

[1] van de Waterbeemd, H.; Gifford, E. ADMET *in silico* Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204. DOI: 10.1038/nrd1032.

[2] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M. *et al*. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids R.* **2012**, *40*, D1100–D1107. DOI: 10.1093/nar/gkr777.

[3] Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013**, *53*, 2499–2505. DOI: 10.1021/ci400099q.

[4] Tharatipyakul, A.; Numnark, S.; Wichadakul, D.; Ingsriswang, S. ChemEx: Information Extraction System for Chemical Data Curation. *BMC Bioinformatics* **2012**, *13*, S9. DOI: 10.1186/1471-2105-13-S17-S9.

[5] Minnich, A. J.; McLoughlin, K.; Tse, M.; Deng, J.; Weber, A. *et al*. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 1955–1968. DOI: 10.1021/acs.jcim.9b01053.

[6] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T. etc. KNIME: the konstanz information miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, L. Eds.; Springer New York, 2007; pp 319–326

[7] Gadaleta, D.; Lombardo, A.; Toma, C.; Benfenati, E. A New Semi-Automated Workflow for Chemical Data Retrieval and Quality Checking for Modeling Applications. *J. Cheminformatics* **2018**, *10*, 60. DOI: 10.1186/s13321-018-0315-6.

[8] Williamson, B.; Wilson, C.; Dagnell, G.; Riley, R. J. Harmonised High Throughput Microsomal Stability Assay. *J. Pharmacol. Toxicol. Methods* **2017**, *84*, 31–36. DOI: 10.1016/j.vascn.2016.10.006.

[9] Kim, M. T.; Wang, W.; Sedykh, A.; Zhu, H. Curating and Preparing High-Throughput Screening Data for Quantitative Structure-Activity Relationship Modeling. In *High-Throughput Screening Assays in Toxicology*; Zhu, H.; Xia, M. Eds.; Springer New York, 2016; pp 161-172.

[10] Papadatos, G.; van Westen, G. J.; Croset, S.; Santos, R.; Trubian, S. *et al*. Overington. A Document Classifier for Medicinal Chemistry Publications Trained on the ChEMBL Corpus. *J. Cheminformatics* **2014**, *6*, 40. DOI: 10.1186/s13321-014-0040-8.

[11] Meng, J.; Chen, P.; Wahib, M.; Yang, M.; Zheng, L. *et al*. Boosting the Predictive Performance with Aqueous Solubility Dataset Curation. *Sci. Data* **2022**, *9*, 71. DOI: ;10.1038/s41597-022-01154-3.

[12] Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds. *Sci. Data* **2019**, *6*, 143. DOI: 10.1038/s41597-019-0151-1.

[13] Esaki, T.; Watanabe, R.; Kawashima, H.; Ohashi, R.; Natsume-Kitatani, Y. *et al.* Data Curation can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance. *Mol. Informatics* **2019**, *38*, 1800086. DOI: 10.1002/minf.201800086.

[14] Hunter, F. M. I.; Bento, A. P.; Bosc, N.; Gaulton, A.; Hersey, A. *et al.* Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs. *Chem. Res. Toxicol.* **2021**, *34*, 385–395. DOI: 10.1021/acs.chemrestox.0c00296.

[15] Grzegorzewski, J.; Brandhorst, J.; Green, K.; Eleftheriadou, D.; Duport, Y. *et al*. PK-DB: Pharmacokinetics Database for Individualized and Stratified Computational Modeling. *Nucleic Acids Res.* **2021**, *49*, D1358–D1364. DOI: 10.1093/nar/gkaa990.