COMMENTARY

The Advantages and Challenges of Using Real-World Data for Patient Care

Yunn-Fang Ho[1], Fu-Chang Hu[2,3] and Ping-Ing Lee[4,*]

Clinical studies and real-world data (RWD) are indispensable for continued advancement of patient care and biomedical sciences. The significance and pros/cons of RWD sources, efficacy- or safety-associated intangible factors are identified, and methodologies for properly performing RWD research are discussed. Additionally, multidisciplinary teams that integrate domain knowledge, statistics, and computing engineering are emphasized for practicing investigations of drug-associated factors of intricate pharmacotherapy that is required to attain the aim of precision therapeutics.

RWD in the health field is invaluable to acquire complementary yet indispensable evidence to preclinical and clinical studies. For instance, common acute adverse drug reactions (ADRs) of vaccinations, such as fever, malaise, and local reactions at injection sites, are usually recorded in clinical trials, but the discoveries of rare or late vaccine-related adverse reactions and unintended interactions with other factors must rely on prudent observations in real clinical settings after widespread vaccinations. As an example, intussusception risk of either monovalent or pentavalent oral rotavirus vaccines was not fully captured until the disclosure by a meta-analysis of five postlicensure studies on RWD collected from active and passive surveillance.[1] Additionally, the extent of protective effects by herd immunity can hardly be answered during the development stage of vaccines. However, this crucial benefit of vaccination can be revealed through investigations of RWD, as exemplified in vaccinations for influenza, pneumococcus, and human papillomavirus worldwide.

Like a double-edged sword, pharmaceuticals may treat or trigger diseases, depending on whether they are used judiciously. Pharmacological effects and notable predictors are explored and examined during drug discovery and development to be readily applicable to clinical settings. However, multiple factors affect therapeutic effects, such as drug properties (e.g., physicochemical, pharmaceutic, pharmacokinetic, and pharmacological characteristics), personal attributes (e.g., age, sex, stature, genetics, disease status, and comorbidities), selected regimens (e.g., drug choice, daily/cumulative dose, frequency, duration, and concomitant medications), and healthcare delivery processes. These complex or inexplicable elements associated with drug efficacy and safety, designated as intangible factors, usually require long-term real-world experiences and strenuous research efforts to identify and report.

Further, polypharmacy and drug interactions are intrinsically risky aspects of pharmacotherapy. Nonetheless, solid evidence for elusive intangible drug covariates associated with either therapeutic or toxic effects are often not readily available upon regulatory approval. It is valuable to perform studies with RWD from real settings, as demonstrated in **Table 1**,[2–6] to supplement awaited determinants of drug effects for clinical decision making. By health record (electronic and paper) review,[2] nephrotoxic polypharmacy exhibited a significant association with contrast medium–induced nephropathy among inpatients undergoing contrast-enhanced computed tomography. Through investigations of health records[3] and postmarketing surveillance data,[6] drug dispositions (interacting drugs, amiodarone cumulative dose, and adjusted average daily dose) and host factors (shorter height and smaller body surface area) were reported to be important predictors of amiodarone-related liver injury. Studies of health claims helped identify the inverse associations between dose intensity (bisphosphonates[4] or statins[5]) and the likelihood of unintended risk (esophageal cancer or poststroke epilepsy). These intangible factors of drug effects would be difficult to recognize if without RWD.

Teamwork and in-depth understanding of data sources are keys to successful RWD mining. Collaborations among physicians, pharmacists, pharmacologists, pharmacovigilance specialists, and statisticians would facilitate multiprofessional brainstorming and ensure insightful health data research. A systematic understanding of the strength/weakness and structures/contents of respective data sources is an essential prerequisite to optimal and wise use of RWD. The pros and cons of health records, health claims, and ADR registries and the multiprofessional nature of data science are discussed herein.

First, the quantity and quality of medical information on health records in any large-scale medical center may be sufficiently thorough to dissect issues pertaining to multiple sectors. Health records usually contain information about patient demographics, health habits, comorbidities, laboratory tests and diagnostic examinations, clinical status, and

[1]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan; [2]International-Harvard Statistical Consulting Company, Taipei, Taiwan; [3]Graduate Institute of Clinical Medicine and School of Nursing, College of Medicine, National Taiwan University, Taipei, Taiwan; [4]Department of Pediatrics, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan. *Correspondence: Ping-Ing Lee (pinging@ntu.edu.tw)
Received: June 11, 2019; accepted: July 10, 2019. doi:10.1111/cts.12683

Table 1 Sources and features of RWD to identify complex drug factors

Data source	Health records[2,3]	Health claims[4,5] (e.g., Taiwan National Health Insurance Research Database)	ADR reports[6] (e.g., Taiwan National ADR Reporting System)
1. Features/ strengths	• Institution-level data • Detailed health records (electronic and paper), consisting of patient demographics and clinical characteristics (e.g., height/weight, genetics, allergy/family/social/history, organ functions, and disease status), inclusive details of concomitant diseases, medications, and prescriptions • Full assessment of each individual patient is feasible	• Population-based data • Opportunity for decade-long follow-up period • Efficient data analysis possible because of its structured *secondary* database nature • Sample size large enough to perform subgroup analyses	• Nation-level data • ADRs reported voluntarily across healthcare professionals, pharmaceutical delegates, and the public • The voluminous reports of longitudinal nature facilitate the extraction of knowledge from rare signals
2. Challenges and limitations	• Laborious data collection: individual health records (electronic and paper) review or assessment is essential • Transformation of unstructured data (verbatim—e.g., admission/progress/nursing notes, and diagnostic reports) into structured/schematized ones might be necessary • Generalizability is confined by limited study duration, setting/site, target cohort, and single institution (tertiary care)	• Inadequate clinical information • Unavailable laboratory and body image data • Lack of a patient's genetics, socioeconomic status, health habits, and lifestyle • Drug prescriptions and dispensing records may not fully reflect the real-life drug adherence of the patients	• Possible underreporting, biased reporting or misclassification, incomplete or missing data, stimulated reporting (by regulatory actions or publicity), and duplicate reporting (e.g., from healthcare institutes and the pharmaceutical industry) • Absence of information on population exposure, patients' clinical details, and confirming rechallenge data
3. Data sectors utilized	• Patient demographic data • Comprehensive medical records (outpatient, emergency department, inpatient) • Precise laboratory data files • Complete pharmacy (hospital) dispensing datasets	• Patient registration data sets • Medical data sets (outpatients, emergency department, inpatients) • Pharmacy data sets (hospital, community)	• Constructed data: origin of the report, patient demographics, prescription and comedication details, and ADR type and seriousness • Verbatim: specifics on clinical presentations, comorbid medical conditions, liver biochemistry values, and ADR manifestations and consequences
4. Drug-associated covariates identified	• Nephrotoxic polypharmacy: significant association between four-nephrotoxic polypharmacy and contrast medium–induced nephropathy among inpatients undergoing contrast-enhanced computed tomography • Interacting drugs: drug-related factors (amiodarone cumulative dose, interacting drugs) are significant predictors of amiodarone-associated acute liver injury	• Dose intensity (usage duration, exposure frequency) of bisphosphonates: inversely associated with esophageal cancer risk • Cumulative doses of poststroke statin exposure: inversely associated with PSE risk • Comedications: potential predictors of or protectors against PSE identified	Risk of amiodarone-related liver injury are associated with: • Drug disposition (adjusted average daily dose) • Host factors (shorter height and smaller body surface area)

ADR, adverse drug reaction; PSE, poststroke epilepsy; RWD, real-world data.

prescription and comedication details.[2,3] Although the electronic health record system has become a modern norm, certain unstructured data or narrative verbatim of the written notes in health records still need laborious efforts to be transformed or coded for research purposes if without adequate support of natural language–processing technology. These institution-level routinely documented records, augmented with purposely collected data for specific study aims, can be scrutinized and analyzed to answer scientific questions to assure the quality of health care. Nonetheless, data integration across institutions is not possible if without patient consent and interinstitution agreement.

Second, the health claims (e.g., Taiwan National Health Insurance Research Database; US Surveillance, Epidemiology, and End Results; and UK General Practice Research Database) as an important source of secondary data are usually population based, longitudinally collected, and structurally constructed. The huge sizes of such data facilitate efficient data analyses and make certain subgroup evaluations possible.[5,6] Unfortunately, health claims have inherent limitations, such as discrepancy between claims data and real-life behaviors (e.g., true adherence), inadequate clinical information (e.g., proximity of surrogates to actual disease status), uncertainty in causality, lack of laboratory and body image data, and potentially inconsistent data quality between tertiary and primary care practices. Rational design with sound methodology is certainly critical to the rigor of claims-data studies. Methodic study flow (e.g., appropriate inclusion and exclusion criteria), objective outcome measures, and serial sensitivity analyses are surely required.[5] Third, postmarket spontaneous reporting of ADRs has the merit of identifying effectively new or rare signals.[6] However, the aforementioned limitations of health claims are also applicable to studies using ADR registries. Furthermore, underreporting, reporting biases, misclassification, and incomplete or missing data are also frequently observed.

Translating RWD into scientific evidence is of practicability and implication if the data and research methodology are of acceptable quality with a credible volume.[7] Owing to the continued advancement of computing power over the past decades, analytical techniques are now available to tackle the challenges of managing and analyzing big data. More importantly, a multidisciplinary realm called *data science* has emerged because of dynamic interactions among the experts from three major areas—domain knowledge (e.g., biomedicine), statistics, and computing engineering. Then, with the support of data scientists, the value of humongous RWD in the health field has risen currently compared with dedicated clinical trials.

Statistically, there are two types of data—*observational* and *experimental*. They differ in whether treatment interventions can be implemented and/or randomized. The merit of observational studies with RWD is to achieve high external validity (i.e., generalizability), whereas randomized clinical trials pursue primarily high internal validity. RWD is often criticized by its lack of prospectively collected data through well-designed studies. It is no wonder that studies using RWD are commonly associated with annoying data problems, including faked values, missing data, lack of important covariates, and nonrandom sampling, as well as the aforementioned limitations for studies on health records, health claims, and ADR registries (see **Table 1**), which have raised concerns over the quality of observational studies with RWD and have stirred up many debates about their usefulness.

To avoid traps and slips, numerous facets need to be considered to perform a sound analysis of RWD. Data linkage, checking, and cleaning should be exercised meticulously from the very beginning. Multiple imputations are exercised, where appropriate, for missing data. In particular, to circumvent the various limitations of RWD, researchers are advised to properly employ *post hoc* study designs (i.e., study designs made after having RWD) in preparing working data from *raw* RWD. As an analogy, a cook picks his/her intended list of vegetables from a farm (i.e., RWD) based on the cook's menu design (i.e., study objective/design). In brief, statistical analysis of RWD involves two steps: (i) to make a *post hoc* study design (e.g., a case-cohort study) to prepare a working data set from RWD suitably; (ii) to analyze the working data set pertinently to uncover as much information as possible toward the study goal (e.g., propensity score analysis and Cox's model with time-dependent covariates). The representative flow of an RWD study is illustrated in **Figure 1**. Above all, appropriateness in research design, statistical methodology, outcome prediction, and result validation are all pivotal in dictating the robustness and credibility of RWD studies.

The paradigm of drug discovery and development has gradually been shifted from a linear process to an integrated (circular or cyclic) approach by linking reciprocally earlier stages of discovery to later stages of clinical development and postlicensure practice evidence.[8] Likewise, drug-associated intangible factors relevant to clinical therapeutics could also be collectively uncovered from the beginning of drug discovery to preclinical assessment, various phases of clinical trials, and RWD data from healthcare services (e.g., health records,[2,3] administrative health claims,[4,5] and disease registry), and postmarketing surveillance (e.g., ADR reporting[6]).

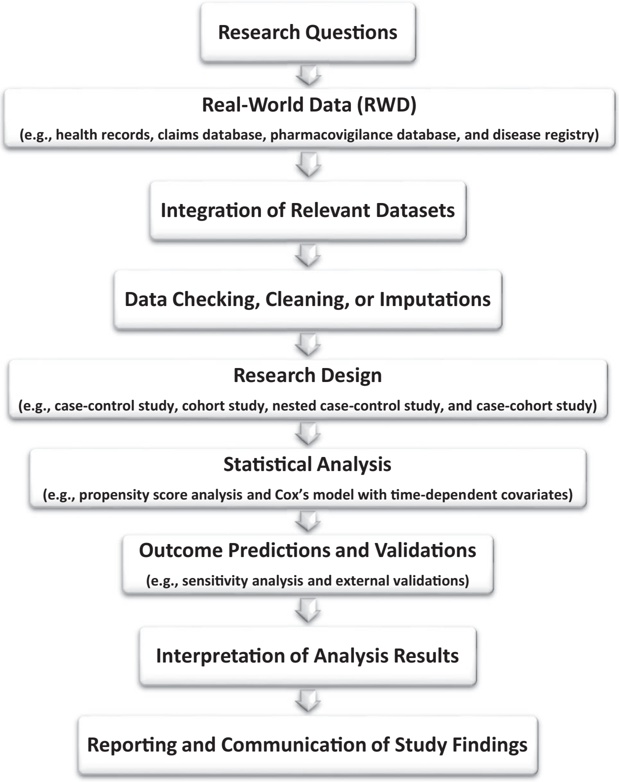
Research Questions
Real-World Data (RWD)
(e.g., health records, claims database, pharmacovigilance database, and disease registry)
Integration of Relevant Datasets
Data Checking, Cleaning, or Imputations
Research Design
(e.g., case-control study, cohort study, nested case-control study, and case-cohort study)
Statistical Analysis
(e.g., propensity score analysis and Cox's model with time-dependent covariates)
Outcome Predictions and Validations
(e.g., sensitivity analysis and external validations)
Interpretation of Analysis Results
Reporting and Communication of Study Findings

Figure 1 Study scheme for real-world data investigations.

Precision therapeutics is an ideal in patient-centric care if emerging intangible factors and conventional concerns (e.g., drug property, personal attribute, and drug regimen) are all sensibly contemplated when prescribing. The goal of precision therapeutics can be ultimately achieved when all drug-associated factors in relation to intricate pharmacotherapy are adequately discerned through prelicensure and postlicensure studies, by probing various types of RWD to the fullest. It is absolutely crucial that the abstraction and interpretation of RWD are exercised vigilantly to avoid unintended mistakes, such as the infamous Google Flu parable, an overestimation of the 2013 peak flu levels by inadequate algorithm in crowdsourcing.[9]

The tide of science and technology is flowing inexorably in favor of data sciences, such as artificial intelligence, big data, and machine learning.[10] Interprofessional collaborations and knowledge exchanges among multiple specialties shall transcend traditional study stereotypes and achieve breakthroughs in effectively utilizing large-scale RWD. Learning from routinely collected healthcare data as much as possible to inform clinical decisions to improve the quality of patient care is projected to thrive in the near future.

Funding. No funding was received for this work.

Conflict of Interest. The authors declared no competing interests for this work.

1. Rosillon, D., Buyse, H., Friedland, L.R., Ng, S.P., Velázquez, F.R. & Breuer, T. Risk of intussusception after rotavirus vaccination: meta-analysis of postlicensure studies. *Pediatr. Infect. Dis. J.* **34**, 763–768 (2015).
2. Ho, Y.-F. *et al.* Nephrotoxic polypharmacy and risk of contrast medium-induced nephropathy in hospitalized patients undergoing contrast-enhanced CT. *Am. J. Roentgenol.* **205**, 703–708 (2015).
3. Ho, Y.-F., Chou, H.-Y., Chu, J.-S. & Lee, P.-I. Comedication with interacting drugs predisposes amiodarone users in cardiac and surgical intensive care units to acute liver injury: a retrospective analysis. *Medicine* **97**, e12301 (2018).
4. Ho, Y.-F., Lin, J.-T. & Wu, C.-Y. Oral bisphosphonates and risk of esophageal cancer: a dose-intensity analysis in a nationwide population. *Cancer Epidemiol. Biomarkers Prev.* **21**, 993–995 (2012).
5. Lin, F.-J., Lin, H.-W. & Ho, Y.-F. Effect of statin intensity on the risk of epilepsy after ischaemic stroke: real-world evidence from population-based health claims. *CNS Drugs* **32**, 367–376 (2018).
6. Ye, J.-H. *et al.* Trends in reporting drug-associated liver injuries in Taiwan: a focus on amiodarone. *Int. J. Clin. Pharm.* **40**, 911–920 (2018).
7. Miksad, R.A. & Abernethy, A.P. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin. Pharmacol. Ther.* **103**, 202–205 (2018).
8. Ginsburg, G.S. & McCarthy, J.J. Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol.* **19**, 491–496 (2001).
9. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
10. Obermeyer, Z. & Emanuel, E.J. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).