PharmacoEconomics (2017) 35:759–765
DOI 10.1007/s40273-017-0513-5

LEADING ARTICLE

Evaluation of Healthcare Interventions and Big Data: Review of Associated Data Issues

Carl V. Asche[1] · Brian Seal[2] · Kristijan H. Kahler[3] · Elisabeth M. Oehrlein[4] · Meredith Greer Baumgartner[1]

Published online: 4 May 2017

Abstract Although the analysis of 'big data' holds tremendous potential to improve patient care, there remain significant challenges before it can be realized. Accuracy and completeness of data, linkage of disparate data sources, and access to data are areas that require particular focus. This article discusses these areas and shares strategies to promote progress. Improvement in clinical coding, innovative matching methodologies, and investment in data standardization are potential solutions to data validation and linkage problems. Challenges to data access still require significant attention with data ownership, security needs, and costs representing significant barriers to access.

Key Points

Data access challenges such as data ownership, security issues, and costs often serve as barriers to data access.

Careful thought is required to fully realize the potential of 'big data' to draw accurate conclusions.

✉ Carl V. Asche
cva@uic.edu

[1] Center for Outcomes Research, University of Illinois College of Medicine at Peoria, 1 Illini Dr, Peoria, IL 61605, USA

[2] Global Health Outcomes, Takeda, Bedford, MA, USA

[3] Novartis Pharmaceuticals Corporation, 1 Health Plaza, East Hanover, NJ 07936, USA

[4] Pharmaceutical Heath Services Research, University of Maryland, Baltimore, MD, USA

1 Introduction

With the rapid expansion of health information technology, healthcare databases contain increasingly complex and informative clinical, utilization, and payment information that can be harnessed to assess and identify effective interventions across time and care settings. Available 'big data' includes an ever-growing repository of clinical, genetic, genomic, social, outcome, and claims information [1]. While the future analytic possibilities of combining these disparate sources is staggering, the opportunity to improve patient care through (1) assessment of treatment effectiveness and (2) outcome prediction in close to real-time is upon us [2]. Recognizing this promise, healthcare investigators and stakeholders across the globe continue to face challenges in fully realizing the potential of 'big data'. As unique multidisciplinary partnerships develop across industry, medicine, and research, data issues specific to validation, linkage, and access must be addressed. This article will discuss these areas and share strategies to promote progress in this rich field.

2 The Issue of Data Validation and Linkage

While analyses of secondary data sources expand the possibility for evaluation and attribution of healthcare interventions well beyond the scope of traditional randomized controlled trials (RCTs), researchers must be prepared for the challenges associated with their lack of internal validity. In addition, as the source and purpose of data components vary greatly, lack of complete or linked information remains a key pitfall to data use and application.

In preparation for secondary analyses, it is critical to understand why the data exists and how this colors our interpretation. While medical claims data facilitates payment and reimbursement for services, electronic health record (EHR) information primarily functions to document the provision of medical care. Patient survey information can provide information on patient-reported progress, engagement in care, health outcomes, and social/environmental exposures, while pharmacy data can provide information on medication use and serve as a proxy for adherence. The separate objectives of the growing number of possible sources definitively shape the type and form of variables collected, whether they are qualitative or quantitative and recorded by a primary or secondary source.

While effective linkage of distinct data sources is our ultimate goal, given the complexity of research variables and documentation sources, establishing confidence in each data component's correctness and accuracy becomes paramount. If there are data validation issues or one piece of the data puzzle is 'broken', this can affect the picture's overall interpretation leading to erroneous attributions and, in the larger context, cast doubt on this exciting growing field as a whole. As we work to dive into larger and larger data pools, specific data validity concerns include unavailable data, reporting bias, and unmeasured confounding.

2.1 Missing, Lost, and Unavailable Data

As investigation into health expands beyond traditional variables and the continuum of care becomes more holistic, some data pieces are missing from the researcher's arsenal. At a minimum, we see this with the lack of easily attainable use and outcome information from over-the-counter (OTC) medication, online services, urgent care clinics, and social services. These key data points are not readily available for integration into larger data sets, limiting researchers' assessment of certain interventions and public health outcomes. For example, discounted generic prescription programs can lead to claims not being submitted to insurance companies for adjudication. This discrepancy can negatively influence tracking of adherence reporting and clinical outcomes [3]. Depending on the intervention under investigation, researchers must attempt to account for the propensity of the patient population to engage with these alternative services.

Moreover, even rich data that is actually captured and recorded, such as in EHRs, is not always available for research. This can include key information on patient progress and outcomes stored within narrative text of these records. Efforts to use Natural Language Processing (NLP) applications have progressed to attempt to address this challenge [4]. Other common missing variables of interest include inpatient drug use and mortality outcomes.

In addition to missing variables of interest, databases can miss full groups of patients. For example, if a US researcher is using commercial insurance claims, most individuals over the age of 65 years can be excluded from this data. In the UK, patients receiving specialist care and the outcome of such care can be omitted from the clinical practice research database (CPRD) [5]. Identifying these 'lost' populations and linking to or developing systems that capture their data components will help better ensure a full clinical picture across the care continuum and patient lifespan is created.

Finally, data can be lost due to definitions created by changing policies. This is seen in the Bundled Payments for Care Improvement (BPCI) initiative operated by the Centers for Medicare and Medicaid Services (CMS). Under BPCI, providers are reimbursed for 'episodes of care' rather than individual services provided. Under this model, services from multiple providers, laboratory tests, and drugs may all be reimbursed under one fee, irrespective of the frequency of laboratory tests or drug utilization. The objective of this payment model is to improve care coordination, reduce unnecessary services, and improve patient care. An indirect result of this payment model is that granularity may be lost in claims data, making it more difficult to analyze the cost or comparative effectiveness of individual procedures or therapies. Procedures and therapies likely to be most impacted include those used during inpatient stays since they are traditionally the most expensive. As payment models transition from fee-for-service to alternative payment models, other data sources such as EHRs may need to be relied on to study these situations.

2.2 Reporting Bias

Another interesting concern regarding secondary data is the potential error or bias arising from the generation, recording, and storage of data within these sources. Terris et al. [6] provide a conceptual framework for sources of bias impacting health state characteristics. This model identifies six spheres where bias can arise, beginning with the patient's propensity to access services. From this point, additional areas of bias can include the physician's propensity to detect, treat, and record, community and system-based factors, and factors associated with the processing and storage of information within the secondary database itself [6]. It then becomes the researchers' and associated organizations' responsibility to understand how this framework applies to the intervention under investigation and identify potential sources of bias. This can be done retrospectively, but in an era of continuous

improvement through data, can also be done prospectively to understand, address, and work to eliminate the systemic layers of bias within the data.

2.3 Unmeasured Confounding

Additionally, reviews of the analysis of secondary healthcare databases will often appropriately recognize the need to address unmeasured confounding when interpreting results [2, 7]. Unlike an RCT, this type of analysis can be prone to omitted-variable bias. Understanding and accounting for these missing variables is key. Current strategies to address confounding include multiple imputation [8], propensity score calibration [9], and use of external validation data [10]. Further, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Task Force has identified the use of stratification analysis, multivariable regression including instrumental variables, and structural modeling techniques to draw valid causal inferences in nonrandomized analysis [11]. In addition, the impact of unmeasured confounding may be quantifiable through sensitivity analyses with methods such as the Rule Out approach or Bayesian modeling with non-informative priors. Given these multiple available strategies, the onus falls to the researchers and data owners to consistently and methodically apply these measures to reduce bias due to confounding.

3 Strategies to Support Data Validation and Linkage

3.1 Expand Data Definitions

While some variable definition changes reduce granularity, others can introduce standardization ripe for comparison. The implementation of ICD-10 in October 2015 has enabled researchers to learn more precise details about clinical diagnoses and trajectories through administrative databases. In a recent review, Outland et al. [12] provide the example of otitis media, a type of ear infection, to illustrate how ICD-10 will improve researchers' ability to understand individual patients' experiences. In the past, with ICD-9 coding, researchers looking at claims databases would only see that a patient experienced an ear infection, not which ear the infection was in or whether it was a recurring infection in the same ear. In contrast, with ICD-10 coding, claims will document whether the infection occurred in the left or right ear and whether the encounter was for a new or recurring infection. These additional details may help reduce current limitations of claims data; however, several important implementation issues may arise for researchers, especially those using data spanning multiple years. For example, coding errors, in particular during years 2015 and 2016, may be widespread and traditional methods for dealing with bias may need adapting. In addition, tools for accurately mapping ICD-9 to ICD-10 codes will be important for reducing bias in cohort discovery and commonly used comorbidity algorithms. To aide in this transition, researchers such as Boyd et al. have begun creating tools on translating ICD-9 to ICD-10 and vice versa [13].

3.2 Match with Greater Precision

Evolving methodologies allow researchers to improve stringency of matching amongst observed patient populations, thereby better simulating randomization. For example, in conventional matching studies, researchers gauge the effectiveness of matching by comparing summary statistics for pre-intervention covariates with post-matching covariates [14]. However, this approach only compares a limited number of observed characteristics. To reduce this bias, machine learning approaches, such as Linden and Samuels' [14] and Yarnold and Soltysik's [15, 16] use of optimal discriminant analysis (ODA) may be useful. This method allows one to compare whether study groups are distinguishable based on complex covariate distributions [15, 16], where successful matching is identified by the algorithm's failure to identify discriminating characteristics between groups.

3.3 Link Populations

To promote linkage of disparate data variables, stakeholders have invested in creating system-level data warehouses, with varying degrees of success [17]. In this vein, American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) developed a framework promoting linkage, recommending investment in data standardization, transparency, accuracy/completeness of data, encouraging provider participation, financial sustainability, and providing feedback to providers [18, 19]. Key lessons regarding linkage can be learned from large clinical databases and registries across the globe:

- The Clinical Practice Research Database (CPRD) of the UK collects data from nearly 700 general practitioners providing care to over 11.3 million patients [5]. This anonymized medical record database allows practitioners to share patient-level information with multiple, trusted third-party systems and has been used in over 1000 epidemiologic studies worldwide [5].
- The Medical Expenditure Panel Survey (MEPS) is a nationally representative survey of US healthcare

service utilization, associated costs, and the breadth of health insurance amongst American families, individuals, and healthcare providers [20, 21] (https://fcsm.sites.usa.gov/files/2014/05/H2_Mirel_2013FCSM.pdf).
- National healthcare registries across Nordic countries have been harnessed for population-level epidemiological research and highlight the use of country/entity-specific ethics committees (ECs) to approve trans-Nordic research [22].
- The ACS NSQIP was developed by the Veterans Association to collect data and reports risk-adjusted surgical outcomes across all participating hospitals to promote quality improvement. Since 2004, this program has been available to private sector hospitals upon subscription, with over 200 hospitals participating [23].

3.4 Digest of Databases

The examples above highlight how current success in large-scale data inventories has, to date, been more often associated with government-led initiatives and organized as either administrative databases or disease/specialty-specific clinical registries [24]. However, researchers are increasingly interested in linking data sources, patient-generated or crowd-sourced data, and opportunities to conduct research among international populations. To help researchers understand the potential for linkage beyond specific diseases or administrative constructs, ISPOR is currently redesigning their Digest of Databases [25]. In addition to identifying databases worldwide through crowdsourcing, the corresponding Working Group is modifying database questionnaires to reflect emerging sources of data, such as patient-generated databases, genomic, or social media sources, where natural language processing may be employed for analyses. Growing interest in genomic testing will also present data processing challenges. To ensure that the Digest includes accurate information, the Working Group has proposed relying on both data vendors and ISPOR member 'peer reviewers,' who have used the database for their own research, to complete database questionnaires.

4 The Issue of Data Access

Even when data pieces are validated and able to be linked, the question of who has access to this information rises to the forefront of the researcher's agenda. Key components of data access include ownership (i.e., investigator, public/private institution, government, industry, individual), information security and privacy, and cost.

4.1 Data Ownership

Currently, there remains great debate over the ownership of medical, prescription, and billing data and whether this is a private or public commodity [26]. Proponents of public ownership argue that it removes the potential for the formation of data analysis monopolies and the selling of de-identified information for gross private gain. Opponents argue private ownership could instead increase the potential for competition in services and the development of public health research entities who can produce more accurate and complex analyses [26]. Further, regardless of ultimate public or private ownership, the role of the individual healthcare consumer, the original owner, in sharing and providing this information remains extremely complex. For robust data to be available, consumers must continually consent (across access points) to share their information and this relies on a foundation of privacy and data security.

4.2 Security and Privacy

The powerful relationship between ownership, security, and privacy concerns is highlighted by the experience of the National Health Service (NHS) in the UK and the failed implementation of its proposed care data initiative in 2015. This initiative aimed to enable large-scale data sharing across the NHS records with both translational healthcare researchers and businesses, including insurers [27]. While the program held significant promise in assessing and improving nationwide health outcomes, its failure was in part due to (1) the lack of public trust in the ensured confidentiality of their private medical information and (2) the lack of transparency in who would have access to this data [27]. This concern of inadequate de-identification of data poses a significant barrier to many such current and future large-scale initiatives. While implementing clear, anonymization methodologies helps address this barrier, as the number of varying, unregulated data sources expands and data variables increase, the potential for and ease of re-identification increases. Further, it becomes necessary to establish the onus of who is responsible for enforcing these methodologies as data changes hands. Beginning in 2000, in the US, healthcare providers and insurance plans have used the Safe Harbor method of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule to anonymize and de-identify data. This method removes 18 unique patient identifiers (16 identifiers for healthcare operations, research, or public health purposes) from stored, shared records. The sufficiency of this de-identification and lack of accountability for re-identification has been of public concern since its inception [28]. For instance, there are specific concerns that certain types of information are at higher risk for re-identification, not

encompassed by Safe Harbor, including genetic information, diagnosis codes, longitudinal data, and free-form entries [28, 29]. While the actual rate of recorded re-identification of data is extremely low, the public's perception or mistrust of data usage remains a major concern. In response to this concern, the Center for Democracy and Technology (CDT) proposed four policies to establish trust in de-identified data: (1) prohibiting unauthorized re-identification of de-identified data; (2) ensuring a robust de-identification methodology; (3) establishing reasonable security safeguards; and (4) increasing public transparency in data use [28, 30].

Emerging partnerships between public and private organizations may improve researchers' ability to access data and thereby control for additional previously unobserved confounders or missing data in analyses. However, as complexities arise, so too will the costs associated with analyzing the data.

4.3 Cost and Availability

Alongside safety and privacy concerns, the resources involved in linking data sets can often pose an unsurpassable barrier for many researchers. The resources include the cost of the data itself, investigator time, and the additional investment in compliant technological equipment (i.e., in the US, research systems must meet HIPAA standards) capable of processing large amounts of data [31]. Even if these constraints were not an issue, the question becomes what information is available for secondary research and possible acquisition? A huge number of small-scale databases containing pertinent health information are available across the globe. If these more local research efforts and data sets could be more systematically identified and housed, researchers would gain more comprehensive access to information and have an opportunity to formulate and answer more far-reaching, global research questions.

5 Strategies to Support Data Access

5.1 Campaign for Public Support and Understanding

Public campaigns, such as former US Vice President Biden's Cancer Moonshot (https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative), may be helpful in highlighting to the public the huge value potential and relatively small risks associated with sharing data. Additionally, partnerships between patient groups, academic institutions, and other healthcare stakeholders may be an effective means for gaining public support for data collection and sharing. Recently, the Patient-Centered Outcomes Research Institute funded PCORnet to build twenty 'Patient-Powered Research Networks' (http://www.pcornet.org/patient-powered-research-networks/). Through multi-stakeholder partnerships, these networks are able to collect data from hospitals, doctors' offices, and community clinics and are intended for research. For example, the Health eHeart Alliance is a partnership between the University of California, San Francisco, the American Heart Association, and StopAfib.org, among others, and has already recruited over 75,000 patients willing to contribute their data (https://www.health-eheartstudy.org/community).

5.2 Sustainably Fund Research

5.2.1 Risk Assessment Plans

At the point of care, one strategy to build transparency and trust among healthcare consumers suggested by Filkins et al. is to develop privacy risk assessment plans aligned with informed consent materials. These plans could include who will have access to the patient's data, the degree of probability of identification of the data, a clear statement of the assessed risk and a clear description of the established safeguards to anonymize data [32]. Alongside these plans, proper education of healthcare consumers of the personal and public health benefits of shared data should be emphasized.

5.3 Establish Distributed Data Networks

On a larger scale, distributed data networks are currently being used with relative success in both comparative effectiveness and pharmacoepidemiologic research to address issues associated with data access [33]. These networks can be comprised of the diverse group of owners of healthcare data who, maintaining control over their protected data and its use, agree to a common data model [34] with universal specifications with regards to uniform data elements.

6 Conclusion

As 'big data' becomes increasingly available to researchers, opportunities to accurately assess treatment effectiveness and outcome prediction amongst subpopulations never studied in randomized trials are an exciting development. However, to fully realize the potential of these data sources and draw accurate conclusions, careful thought is necessary. In many cases, existing methods can reduce biases caused by data validation challenges, such as unmeasured

confounding or missing data. Meanwhile, solutions to data access challenges are still in flux, with data ownership, security needs, and cost often serving as barriers to data access.

Acknowledgements We would like to express thanks to Marie McWhirter from the University of Illinois, College of Medicine at Peoria for her administrative assistance. All authors contributed significantly to the drafting and revision of the manuscript and approved the final version.

Compliance with Ethical Standards

Conflict of interest Carl Asche, Brian Seal, Kristijan Kahler, Elisabeth Oehrlein, and Meredith Baumgartner have no conflicts of interest to declare.

Funding There was no funding provided for this manuscript.

References

1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Project Hope). 2014;33(7):1123–31 **(Epub 2014/07/10. Eng)**.
2. Schneeweiss S. Learning from big health care data. N Engl J Med. 2014;370(23):2161–3 **(Epub 2014/06/05. Eng)**.
3. Tungol A, Starner CI, Gunderson BW, Schafer JA, Qiu Y, Gleason PP. Generic drug discount programs: are prescriptions being submitted for pharmacy benefit adjudication? J Manag Care Pharm JMCP. 2012;18(9):690–700 **(Epub 2012/12/05. Eng)**.
4. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72 **(Epub 2009/08/18. Eng)**.
5. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36 **(Epub 2015/06/08. Eng)**.
6. Terris DD, Litaker DG, Koroukian SM. Health state information derived from secondary databases is affected by multiple sources of bias. J Clin Epidemiol. 2007;60(7):734–41 **(Epub 2007/06/19. Eng)**.
7. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010;48(6 Suppl):S114–20 **(Epub 2010/05/18. Eng)**.
8. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed). 2009;338:b2393 **(Epub 2009/07/01. Eng)**.
9. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005;162(3):279–89 **(Epub 2005/07/01. Eng)**.
10. Sturmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. Med Care. 2007;45(10 Supl 2):S158–65 **(Epub 2007/10/25. Eng)**.
11. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part III. Value Health J Int Soc Pharmacoecon Outcomes Res. 2009;12(8):1062–73 **(Epub 2009/10/02. Eng)**.
12. Outland B, Newman MM, William MJ. Health policy basics: implementation of the international classification of disease, 10th revision. Ann Intern Med. 2015;163:554–6. doi:10.7326/M15-1933.
13. Boyd AD, 'John' Li J, Kenost C, et al. Metrics and tools for consistent cohort discovery and financial analyses post-transition to ICD-10-CM. J Am Med Inform Assoc. 2015;22(3):730–7. doi:10.1093/jamia/ocu003.
14. Linden A, Samuels SJ. Using balance statistics to determine the optimal number of controls in matching studies. J Eval Clin Pract. 2013;19(5):968–75 **(Epub 2013/08/06. Eng)**.
15. Yarnold P, Soltysik RC. Optimal data analysis: a Guidebook with Software for Windows. Washington, DC: APA Books; 2005.
16. Yarnold P, Soltysik RC. Maximizing predictive accuracy. Chicago: ODA Books; 2016.
17. Lyman JA, Scully K, Harrison JH Jr. The development of health care data warehouses to support data mining. Clin Lab Med. 2008;28(1):55–71 **(Epub 2008/01/16. Eng)**.
18. Klaiman T, Pracilio V, Kimberly L, Cecil K, Legnini M. Leveraging effective clinical registries to advance medical care quality and transparency. Popul Health Manag. 2014;17(2):127–33 **(Epub 2013/10/25. Eng)**.
19. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville. 2014.
20. Services UDoHH. Medical Expenditure Panel Survey. 2017. https://meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp. Accessed March 24, 2017.
21. Mirel LM, SR. Enhancing the Medical Expenditure Panel Survey through Data Linkages. https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/H2_Mirel_2013FCSM.pdf Rockville, MD, 2014. Accessed March 24, 2017.
22. Ludvigsson JF, Haberg SE, Knudsen GP, Lafolie P, Zoega H, Sarkkola C, et al. Ethical aspects of registry-based research in the Nordic countries. Clin Epidemiol. 2015;7:491–508 **(Epub 2015/12/10. Eng)**.
23. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg. 2009;250(3):363–76 **(Epub 2009/08/01. Eng)**.
24. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg. 2015;102(2):e93–101 **(Epub 2015/01/30. Eng)**.
25. ISPOR Digest of International Databases Working Group. Uses, Applications and Future Directions of the ISPOR Digest of International Databases. ISPOR 19th Annual European Congress Vienna, Austria. 2016. Available from: https://www.ispor.org/sigs/Digest_SIG-Forum_Vienna_2016.pdf. Updated November 1, 2016.
26. Rodwin MA. The case for public ownership of patient data. JAMA. 2009;302(1):86–8 **(Epub 2009/07/02. Eng)**.
27. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who Owns the Data? Open Data for Healthcare. Front Public Health. 2016;4:7 **(Epub 2016/03/01. Eng)**.
28. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. J Am Med Inform Assoc JAMIA. 2013;20(1):29–34 **(Epub 2012/06/28. Eng)**.
29. El Emam K. Methods for the de-identification of electronic health records for genomic research. Genome Med. 2011;3(4):25 **(Epub 2011/05/06. Eng)**.
30. Blobel B. Paradigm changes of health systems towards ubiquitous, personalized health lead to paradigm changes of the security and privacy ecosystems. Int J Biomed Healthc. 2015;3(1):75–81.

31. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health Serv Res. 2010;45(5 Pt 2):1468–88 **(Epub 2010/11/09. Eng)**.
32. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8(3):1560–80 **(Epub 2016/05/18. Eng)**.
33. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010;48(6 Suppl):S45–51 **(Epub 2010/05/18. Eng)**.
34. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci USA. 2016;113(27):7329–36 **(Epub 2016/06/09. Eng)**.