Are Publicly Funded Health Databases Geographically Detailed and Timely Enough to Support Patient-Centered Outcomes Research?

Soojin Min, PhD[1]*, Laurie T. Martin, PhD*[2]*, Carolyn M. Rutter, PhD*[2]*, and Thomas W. Concannon, PhD*[2,3]

[1]School of Economic, Political and Policy Sciences, The University of Texas at Dallas, Richardson, TX, USA; [2], The RAND Corporation, Santa Monica, CA, USA; [3]Tufts Clinical and Translational Science Institute, Tufts University School of Medicine, Boston, MA, USA.

Emerging health care research paradigms such as comparative effectiveness research (CER), patient-centered outcome research (PCOR), and precision medicine (PM) share one ultimate goal: constructing evidence to provide the right treatment to the right patient at the right time. We argue that to succeed at this goal, it is crucial to have both timely access to individual-level data and fine geographic granularity in the data. Existing data will continue to be an important resource for observational studies as new data sources are developed. We examined widely used publicly funded health databases and population-based survey systems and found four ways they could be improved to better support the new research paradigms: (1) finer and more consistent geographic granularity, (2) more complete geographic coverage of the US population, (3) shorter time from data collection to data release, and (4) improved environments for restricted data access. We believe that existing data sources, if utilized optimally, and newly developed data infrastructures will both play a key role in expanding our insight into what treatments, at what time, work for each patient.

KEY WORDS: PCOR; CER; PM; health database; data utility.

J Gen Intern Med 34(3):467–72
DOI: 10.1007/s11606-018-4673-6

INTRODUCTION

Emerging health care research paradigms, such as comparative effectiveness research (CER), patient-centered outcome research (PCOR), and precision medicine (PM), show promise for addressing some of our most pressing health care challenges. CER aims to allow all relevant stakeholders including patients, caregivers, providers, payers, and policy makers to make well-informed health care decisions based on "evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care."[1] Through stakeholder-engaged research, PCOR particularly emphasizes the engagement of patients.[2, 3] PM evaluates diagnostic, prognostic, and therapeutic approaches precisely tailored to individual patients.[4] All of these research paradigms share one ultimate goal: constructing evidence on the right treatment for the right patient at the right time.[3, 4] For this reason, timely access to individual- or event-level data collected in a timely manner is crucial.

Granular geographic information in data is particularly important in observational studies. Influenced by population characteristics and the availability and accessibility of care, patient and care-provider behaviors can vary significantly over space, resulting in different treatment choices and outcomes by region.[5] If left unaddressed, geographic variation in patient and care-provider behaviors can be a potential source of confounding in observational CER, which often relies on health care utilization data.[5, 6] Individual-level geocoded information is desirable, since spatially aggregated data is susceptible to risk of ecological fallacy.[7] Without granular location information and timely access to data, it is difficult to generate valid, relevant, and timely evidence of comparative effectiveness for different treatments or prescribed medications.

Extensive resources have been invested in establishing data infrastructures to support emerging research frameworks. For example, the Precision Medicine Initiative in 2015 began assembling a national patient-centered longitudinal cohort consisting of more than one million voluntary participants.[8] Additionally, Patient-Centered Outcomes Research Institute (PCORI) operates PCORnet, a distributed research network containing medical information on 128 million Americans. PCORnet is also working with partners to extend its scope to include claims data.[9] New data systems can take substantial time to resolve legal and security challenges before becoming accessible to researchers, however, so existing data continues to be an important resource for CER, PCOR, and PM.

Federal and state governments have long supported the collection of data that can be used for research to improve health care through government agencies and government-affiliated or funded non-profit organizations. This includes population-based health surveys, administrative health data, disease registries, claims data, and clinical trials data. Public use files created using de-identified data are one method for encouraging the use of these data sources.[10] However, public

Prior presentations: This paper has not been presented at any previous conferences.

Received February 7, 2018
Revised June 28, 2018
Accepted September 9, 2018
Published online September 20, 2018

use files are often insufficient to support observational research in the realm of CER, PCOR, and PM. Measures taken to protect personally identifiable information include limited use provisions and data use agreements, secure data environments, and data perturbation (including aggregation).[11, 12]

METHODS

With these changing research demands in mind, we assessed the ability of 20 widely used health data sources to meet the needs of new research paradigms (Tables 1 and 2). All sources were publicly funded, and some included population-based survey systems. Data that meets the HIPAA Privacy Rule's Safe Harbor standards without any restrictions on access is classified as "public data," while data that requires a data use agreement (DUA) is classified as "limited use data." We generalize both of these as "non-restricted data." On the other hand, data accessed only after human subjects committee review and approval and/or via secure data rooms such as federal statistical research data centers (RDCs) are classified as "restricted data." We examined available non-restricted and

Table 1 Spatial Granularity and Access Process of Publicly Funded Health Databases and Population-Based Survey Systems[13–32]

Entity	Health database	Type	Non-restricted data (public or limited use data)		Restricted data	
			Lowest level granularity	Access	Lowest level granularity	Access
CDC	NHIS	Repeated cross-section surveys	Census region (personal file)	Public	Varies by files (county, latitude and longitude, state)*	The RDC approval[‡]
	NHANES	Repeated cross-section surveys	Nation	Public	Latitude and longitude (HUD geocode files)	The RDC approval[‡]
	NAMCS, NHAMCS	Repeated cross-section surveys	Nation	Public	ZIP code	The RDC approval[‡]
	BRFSS	Repeated cross-section surveys	State	Public	Varies by states (ZIP code, county)	Varies by states (committee approval or DUA only)
CMS	MCBS	Short panel studies	Nation ZIP code	Public DUA	–	–
	Medicare Claims	Claims	Nation County	Public (PUF) DUA	ZIP code (research identifiable files)	Privacy board approval
	MAX files	Claims	–	–	ZIP code (research identifiable files)	Privacy board approval
AHRQ	MEPS	Short panel studies	Nation	Public	Census block group	Committee approval[‡]
	CAHPS	Short panel studies	Nation (research dataset)[§]	DUA	–	–
	HCUP	Claims	Varies by states (5- or 3-digit ZIP code)[‖]	DUA	–	–
APCD Council, State Health Dept.	APCDs	Claims	Varies by states (3-digit ZIP code, county, urban/rural)	Public	ZIP code	Committee approval in each state
Add Health, Carolina Population Center	Add Health	Long panel/ cohort studies	Nation	Public	Latitude and longitude (ancillary data)	Restricted-use data contracts
NIA, UMISR	HRS	Long panel/ cohort studies	Nation	Public	ZIP code, census tract	Committee approval
The Urban Institute	HRMS	Long panel/ cohort studies	Nation Census region	Public DUA	–	–
NCI	SEER	Cancer registries	County	DUA	–	–
Entity	Database	Type	Non-restricted data (public or limited use data)		Restricted data	
			Lowest level granularity	Access	Lowest level granularity	Access
Census Bureau	ACS	Repeated cross-section surveys	Census block group	Public	Census block (demographic microdata)	The RDC approval[‡]
	SIPP	Short panel studies	State	Public	Census tract (demographic microdata)	The RDC approval[‡]
Census Bureau /BLS	CPS	Repeated cross-section surveys	County**	Public	Census tract (demographic microdata)	The RDC approval[‡]

(continued on next page)

Table 1. (continued)

Entity	Health database	Type	Non-restricted data (public or limited use data)		Restricted data	
			Lowest level granularity	Access	Lowest level granularity	Access
BLS	NLS	Long panel/ cohort studies	Nation	Public	Varies by files (ZIP code, census tract, latitude and longitude)[††]	The BLS approval[‡]
UMISR	PSID	Long panel/ cohort studies	State	Public	Census block	Committee approval

[*] County (in-house household file), household latitude and longitude (HUD geocode file), state (preliminary quarterly microdata)
[†] FIPS County codes are available but most counties are not identified
[‡] Approved researchers need to be physically present at the facilities within federal data centers to use data
[§] Descriptive information of states may be included
[‖] Nation (nationwide file), 5- or 3-digit ZIP code (state-specific file)
[¶] Spatial granularity is available in some states
[**] FIPS County codes are available but most counties are not identified
[††] ZIP code/census tract (young adult demographic microdata), latitude and longitude (woman demographic microdata)
Abbreviations: CDC, Centers for Disease Control and Prevention; CMS, Centers for Medicare and Medicaid Services; AHRQ, Agency for Healthcare Research and Quality; APCD, All Payer Claims Databases; Add Health, National Longitudinal Study of Adolescent to Adult Health; NIA, National Institute on Aging; UMISR, University of Michigan Institute for Social Research; NCI, National Cancer Institute; BLS, Bureau of Labor Statistics; NHIS, National Health Interview Survey; NHANES, National Health and Nutrition Examination Survey; NAMCS, National Ambulatory Medical Care Survey; NHAMCS, National Hospital Ambulatory Medical Care Survey; BRFSS, Behavioral Risk Factor Surveillance System; MCBS, Medicare Current Beneficiary Survey; MAX, Medicaid Analytic Extract; MEPS, Medical Expenditure Panel Survey; CAHPS, Consumer Assessment of Healthcare Providers and Systems; HCUP, Healthcare Cost and Utilization Project; HRS, Health and Retirement Study; HRMS, Health Reform Monitoring Survey; SEER, Surveillance, Epidemiology, and End Results Program; ACS, American Community Survey; SIPP, Survey of Income and Program Participation; CPS, Current Population Survey; NLS, National Longitudinal Surveys; PSID, Panel Study of Income Dynamics; PUF, Public Use File; HUD, Housing and Urban Development; RDC, Research Data Center

restricted datasets of each database and found four ways current procedures could be improved to better support CER, PCOR, and PM: (1) finer and more consistent geographic data granularity, (2) more complete geographic coverage, (3) shorter time from data collection to data release, and (4) improved environments for restricted data access.

RESULTS

Individual-level geographic information and consistent geographic granularity within a database can be beneficial for observational CER involving geographical variation in treatments and outcomes. Throughout the datasets we examined, we observed substantially different levels of geographic granularity. In non-restricted datasets, geocoded information was generally not available below the 5-digit ZIP code level. Public data, in particular, generally did not even include the 5-digit ZIP code due to compliance with HIPAA standards. While restricted datasets provided relatively finer geographic data, only 10 of the 15 restricted datasets included geocoded information at the census tract level or below (Table 1). Among those, just four of them provided geographic information at the individual level, such as respondent coordinates or residence addresses. These variations in the spatial granularity of restricted datasets across databases indicate differences in the interpretation and application of HIPPA standards, suggesting that there is potential for improvement—wider release of geographic information for research purposes.

Some of the differences in the availability of geographic information result from differences in data access requirements across states, particularly when a database is a collection of state-based systems. For example, although the Centers for Disease Control and Prevention (CDC) established the Behavioral Risk Factor Surveillance System (BRFSS) with state health departments, spatial granularity continues to vary across state-specific datasets. While at least 12 states release respondent data at the ZIP code level, some states do not release any geocoded information at all. In addition, some states require human subjects committee review prior to release of survey information while others require only a DUA. Meanwhile, the Healthcare Cost and Utilization Project (HCUP), supported by the Agency for Healthcare Research and Quality (AHRQ), distributes its data using a single system with a uniform DUA procedure across states through a federal-state-industry partnership. Despite the use of a common DUA, the geographic granularity varies across state-specific datasets. While most states participating in HCUP provide data with 5-digit ZIP codes, some states release only 3-digit ZIP codes or omit geocode identifiers altogether.[33] It may be possible to standardize geographic granularity across these state-specific datasets through collaboration among states, accompanied by coordination at the federal level.

Second, expanding existing databases to improve geographic coverage would increase their value as resources for CER, PCOR, and PM. In particular, this would help fill gaps in

Table 2 Release Interval of Publicly Funded Health Databases and Population-Based Survey Systems[13–32]

Entity	Health database	Approximate interval between collection and release of data
CDC	NHIS	6 months, 5 months (preliminary quarterly microdata)
	NHANES	9 months–1.5 years
	NAMCS, NHAMCS	2–3.5 years, 7 months–3.5 years (restricted data)
	BRFSS	6–9 months (national data), 9 months–1.5 years (varies by states)
CMS	MCBS	1–1.5 years (limited data sets)
	Medicare Claims	1–1.5 years (limited, RIF data) 4.5 months (RIF quarterly data)
	MAX files	2 years or more
AHRQ	MEPS	1–2 years
	CAHPS	1.5 years
	HCUP	1.5–3.5 years (varies by states)
APCD Council, State Health Dept.	APCDs	3 months–2.5 years (varies by states)
Add Health, Carolina Population Center	Add Health	1 year or more (varies by files)
NIA, UMISR	HRS	1 year (core file), 6 months (core early release)
The Urban Institute	HRMS	9 months–2 years
NCI	SEER	6 months[#]
Entity	Database	Approximate interval between collection and release of data
Census Bureau	ACS	9 months–1 year
	SIPP	6–9 months (core file), 1 year (topical module)
Census Bureau /BLS	CPS	6 months-1 year (public use microdata file),1 year or more (demographic microdata)
BLS	NLS	9 months and more (varies by versions), 1 year (confidential file)
UMISR	PSID	1 month (early release)–2 years

[#] *Data submission is 22 months after the close of a diagnosis year. Data released in April 2017 was submitted in November 2016 containing cases diagnosed through 2014*

knowledge until newly developed databases specifically designed to support these emerging research paradigms come online. For example, the Precision Medicine Initiative Cohort Program is a new resource aimed at supporting research on health and disease outcomes, with extensive, comprehensive, and representative information that will ultimately benefit the entire US population.[34] In databases where states choose whether to participate, national datasets are often geographically incomplete due to missing states. In 2014 state-specific HCUP files, datasets from 29 states were available in the State Inpatient Database (SID), 19 in the State Ambulatory Surgery and Services Database (SASD), and 21 in the State Emergency Department Database (SEDD).[13] The presence of missing states in these databases could undermine their ability to estimate outcomes at a national level, although information may still be useful for guiding state-level policy. All Payer Claims Databases (APCDs), on the other hand, are moving toward more complete coverage information across state-level populations. While APCDs have been established in 16 states, the program continues to expand into other states and includes development of a standardized format for data collection across states.[35] Efforts among states to establish APCDs with standardized data and more complete geographic coverage of the US population indicate promising progress toward better supporting the emerging research paradigms.

Third, as timely access to individual- or event-level data is crucial for addressing emerging research questions, we found room for improvement in the duration from data collection to release. In Table 2, we see that the time from collection and release of data varies substantially across databases from a month to years. While some databases, such as National Health Interview Survey (NHIS), Medicare Claims, and Panel Study of Income Dynamics (PSID), release certain datasets within five months, other databases do not release data until years after collection.

PCORI's strategies emphasize timeliness in comparative research from patient-centered studies to better support informed health care decisions.[36] Time lag in data release may impede researchers' ability to compare new treatments in a timely manner.[37] Advances in health informatics technology are improving the feasibility of more rapid release of data. For example, adopting electronic health records can improve the timeliness of public health surveillance systems, particularly those that are disease related, through near real-time data reporting.[38, 39] Promoting Common Data Elements (CDEs) in clinical databases can be a useful approach as well. As an initiative of the National Institutes of Health (NIH), CDEs allow for standardization of data across clinical studies, enabling the effective sharing of quality data.[40] This is relevant to global FAIR Data Principles, a set of guidelines intent on making data findable, accessible, interoperable, and reusable.[41] Timely data access implies that the information reflects current health outcomes and ultimately supports the goal of new paradigms, which is ensuring that patients get the right treatment at the right time.

Lastly, improving the restricted data access environment can assist researchers in CER, PCOR, and PM with accessing individual-level data in a timely fashion. Since the level of geographic granularity in unrestricted data is limited due to privacy protection measures, researchers are encouraged to utilize restricted data sources such as CDC, AHRQ, and the Census Bureau. The challenge is that, currently, the time and resource costs to gain access to granular restricted data can be significant. Data access often includes human subjects committee review and approval. Data may also require approved researchers to carry out analyses within a secure data room at federal RDCs—an additional risk-mitigation measure to limit matching with outside data sources and re-identification.

DISCUSSION AND CONCLUSION

Inevitably, a fundamental trade-off must be made regarding the balance of risk of harm to individuals through the release of personally identifiable information against the potential public benefit that results from research. The use of secured data portals, such as the CMS Virtual Research Data Center (VRDC), has enormous potential to improve qualified data access and to encourage quality CER and patient-centered observational research. Data portals provide the security of physical data centers, but utilize technology to provide researchers with timely, remote access to data in a secure, efficient, and cost-effective manner. The use of data portals allows maintenance of data safeguards while improving timely data access.

Efforts to address these issues can be broken into short-term and long-term goals. Short-term goals focus on increasing the utility of existing databases to sufficiently support emerging research paradigms until new data infrastructures such as PCORnet or the National Precision Medicine Cohort become fully operable. Long-term goals include adaptation of existing databases to new research paradigms, allowing them to better support emerging health research questions and to continue to complement new data sources. In that respect, further discussion is needed to explore the unique uses of existing health databases in the era of big data and advancing health informatics technology, where near real-time data provision will eventually become the norm.

Publicly funded health databases have for years provided opportunities for researchers to make significant contributions to improving health and health care in the nation. Unprecedented advancements in the information landscape—including big data, wearable technology, real-time informatics, and others—have the potential to completely change the way research in the health field is conducted. While these developments bring an incredible amount of potential for advancement in the ways researchers access health information and develop an understanding of treatment options and health outcomes, they also bring about similarly unique challenges in the way stakeholders collect, disseminate, and analyze health data. Utilizing the expanding availability of information while continuing to protect the trust and confidentiality of individual health information is and will continue to be a challenge. As with the rapidly evolving data landscape, data structures, legal and regulatory procedures, and research methods will need to evolve in order to balance each other effectively. Emerging research paradigms, in alignment with the growth of available data, have the potential to expand this progress even further and to provide even greater insight into what treatments, at what time, work for each patient.

Acknowledgements: *The authors wish to thank Daniel A. Waxman, MD, PhD, for his helpful review and comments on the manuscript.*

Corresponding Author: *Soojin Min, PhD; School of Economic, Political and Policy Sciences, The University of Texas at Dallas, Richardson, TX, USA (e-mail: soojin.min@utdallas.edu).*

Funding *This study was funded by contract HHSM-500-2014-00036I (PD Concannon), titled "Evaluation of From Coverage to Care (C2C)," from the Centers for Medicare and Medicaid Services (CMS). The content of this manuscript does not necessarily reflect the views or policies of CMS. The agency had no role in the collection, analysis and interpretation of the findings, or approval of the finished manuscript.*

Compliance with Ethical Standards:

Conflicts of Interest: *The authors declare that they have no conflict of interest.*

REFERENCES

1. **Ratner R, Eden J, Wolman D, Greenfield S, Sox H.** Initial national priorities for comparative effectiveness research. Institute of Medicine Washington, DC: National Academies Pr.; 2009.
2. **Concannon TW, Fuster M, Saunders T**, et al. A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research. J Gen Intern Med 2014;29(12):1692–1701.
3. **Kronenfeld JJ, Parmet WE, Zezza MA.** Debates on US Health Care. Los Angeles: SAGE; 2012.
4. **Mirnezami R, Nicholson J, Darzi A.** Preparing for precision medicine. N Engl J Med 2012;366(6):489–491.
5. **Root ED, Thomas DS, Campagna EJ, Morrato EH.** Adjusting for geographic variation in observational comparative effectiveness studies: a case study of antipsychotics using state Medicaid data. BMC Health Serv Res 2014;14(1):1.
6. **Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES.** Regional variations in diagnostic practices. N Engl J Med 2010;363(1):45–53.
7. **Openshaw S.** The modifiable areal unit problem. Vol 38. Norwich, England: Geobooks; 1983.
8. **Collins FS, Varmus H.** A new initiative on precision medicine. N Engl J Med 2015;372(9):793–795.
9. PCORnet. PCORnet Data. Available at: http://www.pcornet.org/pcornet-data/. Accessed 10 Oct 2017.
10. **Erdem E, Korda H, Sennett C.** Medicare claims data as public use files: a new tool for public health surveillance. J Public Health Manag Pract 2014;20(4):445–452.
11. **Fefferman NH, O'Neil EA, Naumova EN.** Confidentiality and confidence: is data aggregation a means to achieve both?, J Public Health Policy 2005;26(4):430–449.
12. **Reiter JP, Kinney SK.** Commentary: Sharing confidential data for research purposes: a primer. Epidemiology 2011;22(5):632–635.
13. Healthcare Cost and Utilization Project. Databases. Available at: https://www.hcup-us.ahrq.gov/databases.jsp. Accessed 10 Oct 2017.

14. Centers for Disease Control and Prevention. NHIS Dataset Documentation. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2015/. Accessed 10 Oct 2017.
15. Centers for Disease Control and Prevention. Restricted Data. Available at: https://www.cdc.gov/rdc/b1datatype/dt100.htm. Accessed 10 Oct 2017.
16. Centers for Disease Control and Prevention. NAMCS and NHAMCS Restricted Data Available at the NCHS Research Data Center. Available at: http://www.cdc.gov/nchs/data/ahcd/Availability_of_NAMCS_and_NHAMCS_Restricted_Data.pdf. Accessed 10 Oct 2017.
17. Centers for Disease Control and Prevention. Survey Data and Documentation. Available at: https://www.cdc.gov/brfss/data_documentation/index.htm. Accessed 10 Oct 2017.
18. Centers for Disease Control and Prevention. BRFSS State Information. Available at: https://www.cdc.gov/brfss/state_info/index.htm. Accessed 10 Oct 2017.
19. Centers for Medicare and Medicaid Services. Data Documentation and Codebooks. Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Codebooks.html. Accessed 10 Oct 2017.
20. Research Data Assistance Center. Data File Directory. Available at: https://www.resdac.org/cms-data/file-directory. Accessed 10 Oct 2017.
21. Agency for Healthcare Research and Quality. Download Data Files, Documentation, and Codebooks. Available at: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp. Accessed 10 Oct 2017.
22. Agency for Healthcare Research and Quality. Research Datasets. Available at: http://www.ahrq.gov/cahps/cahps-database/data-research/index.html. Accessed 10 Oct 2017.
23. All-Payer Claims Database Council. State Data Access. Available at: https://www.apcdcouncil.org/state-data-access. Accessed 10 Oct 2017.
24. Add Health and Carolina Population Center. Documentation. Available at: http://www.cpc.unc.edu/projects/addhealth/documentation. Accessed 10 Oct 2017.
25. National Institute on Aging and the Institute for Social Research at the University of Michigan. Documentation. Available at: https://hrs.isr.umich.edu/documentation. Accessed 10 Oct 2017.
26. The Urban Institute. Health Reform Monitoring Survey (HRMS) Series. Available at: https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/547. Accessed 10 Oct 2017.
27. National Cancer Institute. Documentation for the Data Files. Available at: https://seer.cancer.gov/data/documentation.html. Accessed 10 Oct 2017.
28. **Lewis DR**, **Chen HS**, **Midthune DN**, **et al.** Early estimates of SEER cancer incidence for 2012: approaches, opportunities, and cautions for obtaining preliminary estimates of cancer incidence. Cancer 2015;121(12):2053–2062
29. **Davis JC**, **Holly BP**. Regional analysis using Census Bureau microdata at the Center for Economic Studies. Int Reg Sci Rev 2006;29(3):278–296.
30. **Davis JC.** RDC Demographic Data. Available at: https://atlantardc.files.wordpress.com/2016/04/davisdemographicdata.pdf. Accessed 10 Oct 2017.
31. Bureau of Labor Statistics. Accessing Data. Available at: https://www.nlsinfo.org/content/getting-started/accessing-data. Accessed 10 Oct 2017.
32. Institute for Social Research at the University of Michigan. Restricted use. Available at: https://simba.isr.umich.edu/restricted/RestrictedUse.aspx. Accessed 10 Oct 2017.
33. Healthcare Cost and Utilization Project. SASD Database Documentation. Available at: https://www.hcup-us.ahrq.gov/db/state/sasddbdocumentation.jsp. Accessed 10 Oct 2017.
34. **Hudson K**, **Lifton R**, **Patrick-Lake B.** The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, ed. 2015.
35. **Curfman G.** All-Payer Claims Databases After Gobeille. Health Aff. 2017;3.
36. **Selby JV**, **Lipstein SH**. PCORI at 3 years—progress, lessons, and plans. N Engl J Med 2014;370(7):592–5.
37. **Fung V**, **Brand R**, **Newhouse J**, **Hsu J**. Using medicare data for comparative effectiveness research–opportunities and challenges. Am J Manag Care 2011;17(7):488.
38. **Birkhead GS**, **Klompas M**, **Shah NR**. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015;36:345–359.
39. **Calderwood MS**, **Platt R**, **Hou X**, et al. Real-time surveillance for tuberculosis using electronic health record data from an ambulatory practice in eastern Massachusetts. Public Health Rep 2010;125(6):843–850.
40. **Biering-Sørensen F**, **Charlifue S**, **DeVivo MJ**, **Grinnon ST**, **Kleitman N**, **Lu Y**, **et al.** Incorporation of the international spinal cord injury data set elements into the national institute of neurological disorders and stroke common data elements. Spinal Cord 2011;49(1):60
41. **Wilkinson MD**, **Dumontier M**, **Aalbersberg IJ**, **Appleton G**, **Axton M**, **Baak A**, **et al.** The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3.