3-31-2016

Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States

Jalpa A. Doshi
University of Pennsylvania School of Medicine, jdoshi@mail.med.upenn.edu

Franklin B. Hendrick
University of Maryland School of Pharmacy, fhend001@umaryland.edu

Jennifer S. Graff
National Pharmaceutical Council, jgraff@npcnow.org

Bruce C. Stuart
University of Maryland School of Pharmacy, BSTUART@RX.UMARYLAND.EDU

Recommended Citation
Doshi, Jalpa A.; Hendrick, Franklin B.; Graff, Jennifer S.; and Stuart, Bruce C. (2016) "Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 4: Iss. 2, Article 8.
DOI: http://dx.doi.org/10.13063/2327-9214.1204
Available at: http://repository.edm-forum.org/egems/vol4/iss2/8

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States

Abstract

Introduction: High quality research regarding treatment effectiveness, quality, and value is critical for improving the U.S. health care system. Recognition of this has led federal and state officials to better leverage existing data sources such as medical claims and survey data, but access must be balanced with privacy concerns.

Methods: We reviewed and catalogued data access policies for a selection of publicly-funded federal and state datasets to investigate how such policies may be promoting or limiting research activities.

Results: We found significant variation in data access policies across federal agencies and across state agencies, including variation for multiple datasets available from the same agency. We also observed numerous indirect hurdles to use of data, including complex data use application procedures, high user fees, and prolonged wait times for data delivery.

Conclusions: Policy makers and data owners should consider making changes to data access policies to maximize the utility and availability of these valuable resources.

Acknowledgements

This study was funded by the National Pharmaceutical Council.

Keywords

Data access policies, public datasets, all-payer claims datasets, privacy, research

eGEMs
Generating Evidence & Methods to improve patient outcomes

Data, Data Everywhere, but Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly Funded Patient-Level Health Care Data in the United States

Jalpa A. Doshi, PhD;[i] Franklin B. Hendrick;[ii] Jennifer S. Graff, PharmD;[iii] Bruce C. Stuart, PhD[ii]

ABSTRACT

Introduction: High quality research regarding treatment effectiveness, quality, and value is critical for improving the U.S. health care system. Recognition of this has led federal and state officials to better leverage existing data sources such as medical claims and survey data, but access must be balanced with privacy concerns.

Methods: We reviewed and catalogued data access policies for a selection of publicly-funded federal and state datasets to investigate how such policies may be promoting or limiting research activities.

Results: We found significant variation in data access policies across federal agencies and across state agencies, including variation for multiple datasets available from the same agency. We also observed numerous indirect hurdles to use of data, including complex data use application procedures, high user fees, and prolonged wait times for data delivery.

Conclusions: Policy makers and data owners should consider making changes to data access policies to maximize the utility and availability of these valuable resources.

[i]University of Pennsylvania School of Medicine, [ii]University of Maryland School of Pharmacy, [iii]National Pharmaceutical Council

Introduction

The United States health care delivery system is moving to a system that rewards quality, effectiveness, and value. Achieving these goals will require expanded research to inform clinical decision-making, promote evidence-based care at reasonable cost, and evaluate innovative system and payment designs. A bedrock requirement for success in these endeavors is the availability of real-world health data.[1] Recognition of this fact has led federal and state health officials to better utilize existing data sources including medical claims and survey data, and to leverage new data sources such as electronic medical records and patient registries. The federal government's Open Government Initiative has increased availability of government data and seeks to increase transparency, public participation, and collaboration around health care data sources.[2] A number of state governments have followed suit with the creation of "all-payer claims data sets" (APCDs).[3] The National Patient-Centered Clinical Research Network (PCORnet),[4] and The Innovation in Medical Evidence Development and Surveillance (IMEDS) program[5] that is offered by the Reagan-Udall Foundation and the United States Food and Drug Administration (FDA), are just two examples of other efforts underway that combine private and public sector data, creating new sources of information for health services research.

The most common data sets released by federal and state agencies constitute "public use files" (PUFs) that report aggregate health data. These files provide too few data elements or are not linkable to other files that are necessary to conduct the kind of inferential health services research needed to improve the effectiveness of our health care delivery system, however. For example, the Centers for Medicare and Medicaid Services (CMS) recently released a PUF with information on the prescription drugs that individual providers prescribed under the Medicare Part D Prescription Drug Program.[6] The PUF provides utilization data (total number of users and total number of prescriptions filled) and spending data (total drug costs) aggregated at the prescriber, drug name, and generic name levels. This provides an invaluable resource for describing patterns of prescription drug use and spending in the Medicare population at the national, state, or prescriber levels. But this file contains no individual patient-level data nor links to other data sources such as medical claims and thus cannot be used to assess drug adherence, evaluate drug safety, or compare outcomes for different drug treatments.

These types of analyses, including comparative effectiveness research and patient-centered outcomes research, require additional health information that is unavailable in PUFs. First, in order to assess treatment outcomes, data should be available at the individual level, rather than aggregated, to preserve the heterogeneity of individual characteristics, behaviors, and treatment impacts. Second, data sets should be longitudinal in order to distinguish temporal patterns of care, such as whether medication use is associated with delayed disease progression or whether patterns of care have an impact on hospitalization risk over time. Third, data on diagnoses, treatments, and outcomes must be dated—both to accurately describe the order of events (and tease out cause and effect) and to account for possible time-related confounding, such as seasonal variation in disease patterns. Fourth, fine-grained geographic detail regarding where patients live and are treated is essential to assessing contextual factors that affect health outcomes, such as residence in an urban versus rural neighborhood. Finally, observations for every data element must be linkable to specific individuals or providers included in the data set, in order to assess patient and provider characteristics that may have an impact on care. For many research questions, other variables and data links may be necessary.[1]

Volume 4 (2016) Issue Number 2

A fundamental policy dilemma is the need to balance access while maintaining individual privacy and confidentiality. A myriad of federal laws apply to publicly funded health care data. For example, statutes such as the Freedom of Information Act and new initiatives such as Open Government seek to increase transparency, openness, and access to federal records. Other statutes, such as the Privacy Act, Health Insurance Portability and Accountability Act (HIPAA), and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) tend to restrict access to federal data (which include health data). The most specific of these is the Privacy Rule of HIPAA.[7] The Rule restricts access to health care data held by specified entities that function as health care providers or health plans (e.g., CMS, the Veterans Health Administration (VHA), and the National Institutes of Health (NIH) in some circumstances). Other entities not typically considered a covered entity (e.g., Agency for Healthcare Research and Quality), often comply with HIPAA. The Rule defines 18 elements of protected health information (PHI) that could be used to identify an individual (names, geographic identifiers below the state level except for the first three digits of a ZIP code, all elements of dates except year, telephone numbers, social security numbers, etc.). HIPAA permits disclosure without authorization when the information is used for approved uses (e.g., treatment, payment and health care operations, and for public health). When information is "de-identified" by one of two methods (formal determination by a qualified expert or removal of all 18 elements of PHI), the data is not considered PHI and is not subject to HIPAA restrictions. However, when disclosure of PHI is necessary, the researcher must obtain authorization from all research subjects, or if this is not feasible, a waiver of authorization from an institutional review board (IRB) or Privacy Board. The Rule also specifies limited data sets (LDS) as data sets with limited PHI (only dates and postal address information at or above the city name level) that can be released without authorization or documentation of a waiver when the data owner and researcher enter into a data use agreement (DUA).

Several federal agencies, particularly CMS, have developed LDS versions of their files and have also established the term "research identifiable files" (RIFs), for files containing PHI elements beyond what are found in their LDS files. In many cases, however, the sensitivity of the PHI found in LDS files is the same or greater than RIFs, and RIFs appear to match the definition of LDS found in the Rule. Inconsistencies also appear in the policies established by federal agencies to access LDS and RIF files. For example, requests for RIFs require more detailed DUAs and higher standards of data security than LDS, and certain LDS requests require IRB approval. Another difference between the file types is that LDS files are available in "shelf-ready" formats and cannot be individualized for specific patient cohorts or alternative sample sizes. In contrast, RIF samples can be customized and are subject to "minimum essential data" requirements that restrict sample cohorts and data elements to those deemed necessary to answer a specific research question.

Furthermore, some publicly funded RIFs are made available only to researchers in academic and nonprofit organizations and cannot be obtained by researchers associated with or funded by commercial entities. Other data sets can be accessed only by employees or contractors of the data owners. Even when data with PHI elements are released, data owners may limit the kinds of research that can be conducted (e.g., only studies with public health interest). Access may also be indirectly restricted by high data acquisition costs, complicated application procedures and licensing restrictions, delays in data turnaround, and costly data security safeguards.

This article catalogues and evaluates common access issues associated with the most widely sought, publicly funded health care databases. We reviewed 116 data sets currently available from federal agencies and state governments. We summarize our findings relating to 19 of these data sets, followed by discussion and recommendations for optimizing access to the kinds of health data necessary to speed our quest toward a high value health system while maintaining safeguards for individual privacy.

Methods

Our search strategy involved a landscape review of federal health agency websites, relevant published papers, gray literature, and personal contacts in order to generate a master list of individual-level health care data files available from federal and state agencies. Our initial search for federal data sets focused on data available from the following agencies or offices: (1) the CMS; (2) the Agency for Healthcare Research and Quality (AHRQ); (3) the FDA; (4) the Health Resources and Services Administration (HRSA); (5) the NIH; (6) the Substance Abuse and Mental Health Services Administration (SAMHSA); (7) the Centers for Disease Control and Prevention (CDC); (8) the Office of the Assistant Secretary for Planning and Evaluation (ASPE); (9) the Office of the National Coordinator for Health Information Technology (ONC); (10) the Census Bureau; and (11) the VHA. To identify data sets produced by state agencies, we searched reports from the All-Payer Claims Database Council (APCD Council) and the State Health Access Data Assistance Center (SHADAC). A brief description of all 116 data sets (98 federal and 18 state) is available online as Additional Supplemental content.

From this master list, we selected 19 data sets (9 federal and 10 state) for detailed review based on four criteria: (1) data were fully or partially publicly funded, (2) the unit of observation was at the individual rather than the population level, (3) data elements included service dates and geographic detail below the level of the state, and (4) data access policies for external researchers were fully developed. For federal agencies like CMS, which maintains multiple data sets meeting our inclusion criteria, we selected the data sets with the broadest policy relevance (e.g., Medicare and Medicaid data) in order to illustrate the data access policies of that agency. In cases where state APCDs made multiple versions of their files available with varying PHI levels, we restricted our review to versions meeting all four of our selection criteria.

Our analysis included the following points for each data set: name and web address or portal; a brief description of data files; type of file (LDS, RIF, etc.); identification of individual-level data (event dates, ages over 89, and lowest geocode of residence); restrictions on data access; data costs; required documentation and approvals; timeliness of data request approvals; data- storage, -release, and -access policies; and the most recent year of data available. We used the data owner's terminology when describing these elements even when they do not correspond to the exact definitions found in the HIPAA Privacy Rule. For example, for the type of file description, CMS refers to certain files as RIFs when in fact they qualify as LDS files under the rule. Similarly, for the required documentation and approvals, the terminology for DUAs is specific to requests for LDS files but CMS also calls the agreements with researchers to access and use RIFs as DUAs. Information was obtained by examining documentation on each agency's or organization's website and by contacting agency personnel to fill in data gaps. Data sets were determined to include PHI elements based on examination of data dictionaries, confirmation from the data owner, or evidence that the data set is available in a form linked to external

files that include one or more PHI data elements (e.g., linkage with individual-level Medicare claims and enrollment files).

Results

Federal Data Sets

Our initial search identified 98 data sets from the federal agencies listed above. A majority of these data sets did not include individual-level data, event dates for medical services, or geographic detail below the level of the state; hence, they were excluded from further review. Table 1 characterizes the 9 federal data sets selected for detailed review.

These 9 data sets represent five federal agencies and include individuals covered under Medicare, Medicaid, the VHA, and commercial insurance plans. The files include the Chronic Condition Data Warehouse (CCW), Medicaid Analytic eXtract (MAX) files, Medicare Standard Analytical Files (SAFs), FDA's Mini-Sentinel distributed data network, the Veterans Health Information and Technology Architecture (VistA) files (including electronic health record data), and 4 national survey or registry-based data sets that are linked to Medicare enrollment and claims: the Medicare Current Beneficiary Survey (MCBS); the Surveillance, Epidemiology, and End Results (SEER) Medicare Linked Database; the Health and Retirement Study (HRS); and the National Health and Nutrition Examination Survey (NHANES).

As required in our selection criteria, each data set included individual-level health information sufficient to conduct high-quality, inferential health services research studies. Data elements include service dates, birth dates (except SEER–Medicare that includes birth month and year only), and date of death. Seven of the 9 include five-digit residential ZIP codes while 2 (the Medicare SAFs and the SEER–Medicare linked files) report residence at the county level. According to the HIPAA Privacy Rule's definitions, all of these files would be designated LDS. However, as noted above, some federal agencies, in particular CMS, designate these files as RIFs, reserving the LDS terminology for files with less sensitive PHI elements. Occasionally, these distinctions are not clear. For example, the MCBS (designated LDS) contains the same PHI data elements as CCW (RIF).

Most federal agencies permit access to users external to the agency; however, there is significant variation across agencies in access restrictions based on type of research, requestor, and funding source. Some agencies (e.g., CMS and the VHA) only make data available for research that advances the mission of their agencies, though it is unclear how that language is enforced. The VHA, which is the only federal agency in our review with electronic health record data available for research, further restricts access to those with Department of Veterans Affairs (VA) appointments although non–VA employees can obtain temporary VA appointments. While there are no explicitly stated restrictions on type of requestor for CCW and MAX files, CMS has not traditionally allowed use of RIFs for commercial purposes, including use by those deemed to be commercial entities.

On the other hand, other CMS data sets (e.g., Medicare SAFs and MCBS) have no such restrictions and are accessible by commercial entities. The SEER–Medicare linked files also have no explicit restrictions on type of requestor, but the NIH requires a letter explicitly indicating that researchers have freedom to publish all findings. The HRS–Medicare linked data set is explicitly restricted to projects funded by a United States government grant, contract, or foundation. The request policies for the FDA's Mini-Sentinel distributed data network are unique among the federal databases we examined—in that access is restricted to the

Table 1. Characteristics and Access for Selected Federal Data Sets Containing Protected Health Information

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
Centers for Medicare and Medicaid Services (CMS)	Chronic Condition Data Warehouse (CCW) files http://www.resdac.org/cms-data/request/research-identifiable-files	Enrollment information and claims data for Medicare Part A, B, and D including service dates, diagnoses, procedures, charges, and payments, plus functional assessment data (MDS, OASIS), Part D plan characteristics and formulary data, prescriber and dispenser characteristics.	RIF	5-digit ZIP code, service dates, birth dates, ages >89, death dates	Only for research that advances CMS mission	No explicit restrictions	No explicit restrictions, but requestor must be deemed independent of commercial funding source
Centers for Medicare and Medicaid Services (CMS)	Medicaid Analytic eXtract (MAX) files http://www.resdac.org/cms-data/request/research-identifiable-files	Enrollment information and inpatient, long-term care, prescription, and other claims and encounter records for all Medicaid recipients. Includes service dates, diagnoses, procedures, charges, and payments.	RIF	5-digit ZIP code, service dates, birth dates, ages >89, death dates	Only for research that advances CMS mission	No explicit restrictions	No explicit restrictions but requestor must be deemed independent of commercial funding source
Centers for Medicare and Medicaid Services (CMS)	Medicare Standard Analytical Files (SAF) http://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/StandardAnalyticalFiles.html	Enrollment information and claims data for Medicare Part A and B including service dates, diagnoses, procedures, charges, and payments.	LDS	County, service dates, birth dates, ages >89, death dates	Only for research that advances CMS mission	No explicit restrictions	No explicit restrictions
Centers for Medicare and Medicaid Services (CMS)	Medicare Current Beneficiary Survey (MCBS) http://www.resdac.org/cms-data/request/medicare-current-beneficiary-survey	Panel survey of a nationally representative sample of the Medicare population. Survey data on socioeconomic and demographic characteristics, health status and functioning, health care use and expenditures, health insurance coverage, and access to care. MCBS files come linked to Part A, B, and D claims records including service dates, diagnoses, procedures, charges, and payments.	LDS	5-digit ZIP code, service dates, birth dates, ages >89, death dates	Only for research that advances CMS mission	No explicit restrictions	No explicit restrictions

Notes: (DHHS) Department of Health & Human Services; (FFS) fee-for-service; (LDS) limited data set; (MDS) Minimum Data Set; (OASIS) Outcome and Assessment Information Set; (PUF) public use file; (RIF) research identifiable file.
Data are based on information available as of November 2014.
[a]"Type of file" is based on the terminology used by the data owner.

Volume 4 (2016) Issue Number 2

Table 1. Characteristics and Access for Selected Federal Data Sets Containing Protected Health Information (Cont'd)

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
National Institutes of Health (NIH) National Cancer Institute (NCI)	Surveillance, Epidemiology, and End Results (SEER)–Medicare Linked Database http://appliedresearch.cancer.gov/seermedicare/obtain/requests.html	Data from cancer registries linked with Medicare enrollment and claims files. Includes demographic characteristics, cancer site and stage, and cause of death information.	LDS	County, service dates, birth (year and month), death dates	Only for research purposes	No explicit restrictions	Commercial funders must provide letter indicating researcher has freedom to publish.
National Institutes of Health (NIH) National Institute on Aging (NIA)	Health and Retirement Study (HRS) linked with Medicare enrollment and claims http://hrsonline.isr.umich.edu/index.php?p=resdat	Biennial panel survey of Americans over the age of 50. Survey data include demographic and socioeconomic characteristics, labor force participation, retirement, health status, and functioning. Can be linked with restricted data including interview and birth date, Medicare enrollment, and Part A, B, And D claims files.	HRS-Restricted Data (NIA), Medicare enrollment and claims-RIF (CMS)	5-digit ZIP code, service dates, birth dates, ages > 89, death dates	Only for research and statistical purposes (NIA). Only for research that advances CMS's mission (CMS)	Only if affiliated with an institution with a DHHS certified Human Subjects review process	The project must be funded by U.S. government grant or contract or foundation.
Centers for Disease Control and Prevention (CDC)	National Health and Nutrition Examination Survey (NHANES) linked with Medicare enrollment and claims http://www.cdc.gov/nchs/nhanes	NHANES data can be linked with Medicare enrollment and Part A, B, and D claims files.	NHANES-PUF (CDC), Medicare enrollment and claims-RIF (CMS)	5-digit ZIP code, service dates, birth dates, ages > 89, death dates	Research must have public health benefit.	No explicit restrictions	No explicit restrictions
U.S. Food and Drug Administration (FDA)	FDA Mini-Sentinel distributed data network	Mini-Sentinel is a pilot project sponsored by the FDA to create an active surveillance system— the Sentinel System— to monitor the safety of FDA-regulated medical products. Mini-Sentinel uses preexisting electronic health care data from multiple sources. Collaborating institutions provide access to data, as well as scientific and organizational expertise. Data include enrollment, demographic, prescription drug, and medical procedure information.	Not specified	5-digit ZIP code, service dates, birth dates, ages >89, death dates	Limited to Mini-Sentinel's public health purposes (e.g., active surveillance, assessment of the impact of FDA actions). Expanded Sentinel Initiative will include broader research component.	N/A for Mini-Sentinel in the pilot phase. Only FDA and Mini-Sentinel Collaborators have access to data.	N/A for Mini-Sentinel in the pilot phase

Notes: (DHHS) Department of Health & Human Services; (FFS) fee-for-service; (LDS) limited data set; (MDS) Minimum Data Set; (OASIS) Outcome and Assessment Information Set; (PUF) public use file; (RIF) research identifiable file.
Data are based on information available as of November 2014.
[a]"Type of file" is based on the terminology used by the data owner.

Table 1. Characteristics and Access for Selected Federal Data Sets Containing Protected Health Information (Cont'd)

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
Veterans Health Administration (VHA)	Veterans Health Information System and Technology Architecture (VistA)	VA-wide information system built around electronic medical record data relating to veterans' health care, with nearly 160 integrated software modules for clinical care, financial functions, and infrastructure.	Not specified	5-digit ZIP code, service dates, birth dates, death dates	No explicit restrictions	Only VA employees. Non–VA employees can obtain a temporary VA appt. and access data at a VA facility	No explicit restrictions

Notes: (DHHS) Department of Health & Human Services; (FFS) fee-for-service; (LDS) limited data set; (MDS) Minimum Data Set; (OASIS) Outcome and Assessment Information Set; (PUF) public use file; (RIF) research identifiable file.
Data are based on information available as of November 2014.
[a]"Type of file" is based on the terminology used by the data owner.

FDA and its Mini-Sentinel Collaborators and Data Partners—given that it is a pilot program established for public health and not research purposes.

Table 2 characterizes data costs, request processes, timelines for approval, data storage and management requirements, and access policies for each of the 9 federal data sets. There is considerable variation across and even within agencies in terms of how data fees are structured. Using the CCW as an example, the current cost for obtaining a random 5 percent sample of Medicare beneficiary records is approximately $35,000 per year of data at the high end. In contrast, the MCBS annual modules cost just $800.

To obtain LDS and RIFs, all data owners require a project summary, data management plan, and DUA. Four of the data sets additionally require a variable selection worksheet for particular variables (e.g., Part D data and functional assessment files). With the exception of the Medicare SAF files and the MCBS, all other data set requests are required to show additional evidence of internal review board (IRB) approval and to undergo privacy board review. While SEER–Medicare requires IRB approval, there is no privacy board review. The timelines and complexity of data request approvals also vary considerably. For instance, it typically takes 8 to 18 weeks to receive approval for the CCW and MAX files, whereas the Medicare SAF files and MCBS DUA applications are typically turned around in under a month.

Another issue of importance to researchers seeking federal data relates to data maintenance and security requirements. HIPAA Privacy, Security, and Breach Notification Rules dictate how entities receiving PHI must protect individual confidentiality through safeguards on transmission and storage of data, safeguards on data access given to investigators, and provisions for data destruction once the research has been completed. All data sets reviewed have rigorous policies on securing the research environment. Although HIPAA sets the same standards for all LDS files, CMS imposes stricter security measures for its RIFs (e.g., CCW and MAX files) compared to its LDS files (e.g., SAF and MCBS). Others (e.g., NHANES–Medicare linked files and VistA) are available only on-site or with remote

Volume 4 (2016) Issue Number 2

Table 2. Data Request Process, Cost, and Timelines; Data Storage and Access Policies; and Latest Years of Data for Selected Federal Data Sets

NAME OF DATA SET	DATA COST[a]	DOCUMENTS & APPROVALS REQUIRED TO SUCCESSFULLY REQUEST DATA SET						TIMELINE FOR DATA REQUEST APPROVAL	DATA STORAGE AND ACCESS POLICIES	MOST RECENT YEAR OF DATA A VAILABLE
		PROJECT SUMMARY	DMP	DUA	VARIABLE SELECTION WORKSHEET	IRB APPROVAL	PRIVACY BOARD REVIEW			
Chronic Condition Data Warehouse (CCW) files	Fixed fee per file type per year of data depending on number of beneficiary lives requested ($$$); data reuse fee ($$); for virtual access fixed fee per user ($$$)	Yes	Yes	Yes	Yes (for Part D data and functional assessment files)	Yes	Yes	-6 to 18 weeks	(1) Physical data files mailed to researcher (rigorous policies on securing research environment), and (2) Remote access	Medicare files (2012)
Medicaid Analytic eXtract (MAX) files	Fixed fee per file type per year of data depending on number of beneficiary lives requested ($$$); data reuse fee ($$); for virtual access fixed fee per user ($$$)	Yes	Yes	Yes	Yes (for functional assessment files)	Yes	Yes	-6 to 18 weeks	(1) Physical data files mailed to researcher (rigorous policies on securing research environment), and (2) Remote access	MAX files (2010)
Medicare Standard Analytical Files SAF)	Fixed fee per year of data ($$); data reuse fees ($)	Yes	Yes	Yes	No	No	No	- 3 to 4 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	All files (2011)
Medicare Current Beneficiary Survey (MCBS)	Fixed fee per year of data ($)	Yes	Yes	Yes	No	No	No	-3 to 4 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	Access to Care (2012), Cost and Use (2010)
Surveillance, Epidemiology, and End Results (SEER)–Medicare Linked Database	Fixed fee per file (per year depending on type of file) by type and number of cancer sites ($$)	Yes	Yes	Yes	No	Yes	No	-4 to 6 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	SEER cases (2011) with Medicare enrollment and claims (2012)

Notes: (CMS) Centers for Medicare & Medicaid Services; (DMP) data management plan; (DUA) data use agreement; (FFS) fee-for-service; (IRB) institutional review board.
Data are based on information available as of November 2014.
[a]$ denotes typically < $1000; $$ denotes typically >=$1000 and < $10,000; $$$ denotes typically > $10,000.
[b]The requestor must first obtain approval, which can take from under 3 months to as much as over a year, depending on the research environment (e.g., computer without access to internet locked in room versus networked computer). The request must then obtain CMS approval, which can take as long as 6 to 18 weeks.

Table 2. Data Request Process, Cost, and Timelines; Data Storage and Access Policies; and Latest Years of Data for Selected Federal Data Sets (Cont'd)

NAME OF DATA SET	DATA COST[a]	DOCUMENTS & APPROVALS REQUIRED TO SUCCESSFULLY REQUEST DATA SET						TIMELINE FOR DATA REQUEST APPROVAL	DATA STORAGE AND ACCESS POLICIES	MOST RECENT YEAR OF DATA A VAILABLE
		PROJECT SUMMARY	DMP	DUA	VARIABLE SELECTION WORKSHEET	IRB APPROVAL	PRIVACY BOARD REVIEW			
Restricted version of the Health and Retirement Study (HRS) linked with Medicare enrollment and claims	Fixed fee per year ($$).	Yes	Yes	Yes	Yes (for Part D data and assessment files)	Yes	Yes	See Table Note[b]	Physical data files mailed to researcher (rigorous data security policies apply)	Linked with CMS Medicare enrollment and claims files (2012)
Restricted version of the National Health and Nutrition Examination Survey (NHANES) linked with Medicare enrollment and claims	Fixed fee per day for on-site access ($); fixed fee per day for staff-assisted research ($).	Yes	N/A	Yes	Yes	Yes	Yes	-6 to 8 weeks	(1) On-site access, (2) remote access, or 3) staff-assisted research option	1999–2004 NHANES linked through 2007 Medicare enrollment and claims files
FDA Mini-Sentinel distributed data network	No costs.	N/A for public health operations						N/A	Data Partners run queries on their own data and provide aggregate results to the Coordinating Center. Data held outside of the Data Partner's environment must meet rigorous data security policies	2014
Veterans Health Information System and Technology Architecture (VistA)	No fee for local VHA data; fixed fee per file per year for national files ($$)	Yes	N/A	Yes	No	Yes	Yes	At least 4 weeks	(1) On-site access, (2) Remote access	2014

Notes: (CMS) Centers for Medicare & Medicaid Services; (DMP) data management plan; (DUA) data use agreement; (FFS) fee-for-service; (IRB) institutional review board.
Data are based on information available as of November 2014.
[a]$ denotes typically < $1000; $$ denotes typically >=$1000 and < $10,000; $$$ denotes typically > $10,000.
[b]The requestor must first obtain approval, which can take from under 3 months to as much as over a year, depending on the research environment (e.g., computer without access to internet locked in room versus networked computer). The request must then obtain CMS approval, which can take as long as 6 to 18 weeks.

access permissions, whereas other data sets are either physically mailed directly to the researcher or are available through special online links.

The recency of the data available to researchers also varies widely across the files we examined. Medicare-only or Medicare-linked files typically have a lag of two to four years (i.e., 2011–2013 data are currently available). The MAX files have a lengthier lag, with most recent data available from 2010. VistA is the most recent and generally available in real time, at least for local facility data.

State All-Payer Claims Data Sets (APCDs)

Our search identified 11 state health agencies that had publicly funded all-payer claims databases (APCDs) available to researchers: Colorado, Kansas, Maine, Maryland, Massachusetts, Minnesota, New Hampshire, Oregon, Tennessee, Utah, and Vermont. At the time of our review, Utah had only a PUF available and thus was not included.

Table 3 characterizes the 10 publicly funded state APCDs selected for detailed review. State APCDs provide unique opportunities to conduct systemwide studies of health care quality, cost, and outcomes. However, the type of data collected and the insured populations represented vary considerably. While all APCDs collect enrollment, medical, and pharmacy claims, 5 APCDs also collect dental claims. All states except Kansas (which includes only state employee health plan and Medicaid data) collect these data for commercially insured populations. More variability is seen in terms of inclusion of Medicare data across APCDs. Three state APCDs currently do not include Medicare data, though New Hampshire and Tennessee plan to include Medicare data in the future; 2 APCDs include only Medicare Advantage (managed care) data; and 5 APCDs include data on both Medicare Advantage and Medicare fee-for-service enrollees.

All state APCDs we reviewed include PHI such as dates of service and birth dates (4 APCDs either provide birth month and year or year only). In contrast to the federal data sets, only 2 APCDs (Kansas and Vermont) include date of death. The lowest level of geographic identifiers varies considerably, with 4 APCDs that include city name, 4 that provide five-digit ZIP codes, and 3 that include street addresses.

We observed significant heterogeneity in data access policies across state agencies, with differences clustering at two extremes. Four (Kansas, Minnesota, Tennessee, and Vermont) provide no or minimal access to external users, whereas the others produce file versions that are broadly available to external researchers. The most PHI-sensitive file version of the Colorado APCD, which includes street addresses, is restricted to requestors from academic institutions and to health care providers. The Massachusetts APCD is restricted to state government agencies, researchers, providers, and qualified individuals. The less PHI-sensitive file version of the Colorado APCD, and the APCDs available from Maine, New Hampshire, and Oregon, have no explicit restrictions on which entities can request data, but they do place restrictions on the types of analyses that can be conducted (e.g., in the public interest, advancing the state authority's mission, and for research or statistical purposes). Access to data on Medicare fee-for-service enrollees is restricted to state government agencies (e.g., Massachusetts and Maryland APCDs), but is available to the public for the Maine APCD. In any case, access to Medicare data from a state APCD is subject to a signed DUA between the state and the CMS.

Table 3. Characteristics and Restrictions on Access to Selected State Funded All Payer Claims Data Sets (APCDs)

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
Vermont Green Mountain Board	Vermont Healthcare Claims Uniform Reporting and Evaluation System (http://gmcboard.vermont.gov/vhcures)	Enrollment information; medical and pharmacy claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicare (FFS under DUA with CMS, Medicare Advantage, Medicare Part D) • Medicaid (FFS and managed care)	Not specified	Street address, service dates, birth dates, ages > 89, death dates	N/A	Only state govt. agencies or contractors	N/A
Massachusetts Center for Health Information and Analysis	Massachusetts All-Payer Claims Database (http://www.mass.gov/chia/researcher/hcf-data-resources/apcd/acessing-the-apcd/learn-how-to-apply-for-apcd-data.html)	Enrollment information; medical, pharmacy, and dental claims; and provider information from the following: • Commercial payers • Third party administrators • Medicare (FFS under DUA with CMS, Medicare Advantage, Medicare Part D) • Medicaid and MassHealth (FFS and managed care)	Level 2 File	City name, service dates, birth dates (month and year only), ages > 89	Only if purpose serves the public interest. Purpose of Medicare FFS data must fall under the DUA with CMS. Use of Medicaid data must be connected with Medicaid program	Only state govt. agencies, researchers, providers, and qualified individuals. Only state govt. agencies can use Medicare FFS data.	No explicit restrictions
Maine Health Data Organization	Maine All-Payer Claims Database (https://mhdo.maine.gov/claims.htm)	Enrollment information; medical, dental, and pharmacy claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicare (FFS, Medicare Advantage, Medicare Part D) • Medicaid (FFS and managed care)	Not specified	City name, service dates, birth dates, ages > 89	Only for research or statistical purposes. Medicare FFS data must fall under the DUA with CMS	No explicit restrictions	No explicit restrictions

Notes: (CMS) Centers for Medicare & Medicaid Services; (DUA) data use agreement; (FFS) fee-for-service.
Data are based on information available as of November 2014.
[a]States in most cases do not use the research identifiable file and limited data set terminology. The current listings come from the language found on state websites.

Table 3. Characteristics and Restrictions on Access to Selected State Funded All Payer Claims Data Sets (APCDs) (Cont'd)

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
New Hampshire Dept. of Health and Human Services	New Hampshire Comprehensive Health Care Information System (https://nhchis.com/NH/# and http://www.gencourt.state.nh.us/rules/state_agencies/he-w900.html)	Enrollment information; medical, pharmacy, and dental claims; and provider information from the following: • Commercial • Medicaid (managed care, no FFS Medicaid in state) • There are plans to include FFS and Part D from the Medicare program in the future.	Commercial Limited Use Data Set	City name, service dates, birth dates, ages > 89	Only for research purposes.	No explicit restrictions	No explicit restrictions
Minnesota Dept. of Health	Minnesota's All-Payer Claims Database (http://www.health.state.mn.us/healthreform/allpayer/)	Enrollment information; medical and pharmacy claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicare (FFS under DUA CMS but the Minnesota Department of Health is also a Qualified Entity, Medicare Advantage, Medicare Part D) • Medicaid (FFS and managed care)	Not specified	City name, service dates, birth dates, ages > 89	N/A	Only state govt. agencies or contractors	N/A
Kansas Dept. of Health and Environment (KDHE) Division of Health Care Finance (DHCF)	Kansas Data Analytic Interface (http://www.kdheks.gov/hcf/medicaid_reports/Health_Care_Market_Reports.html)	Enrollment information; medical, dental, and pharmacy claims from the following: • State Employee Health Plan • Medicaid (managed care, no FFS Medicaid in state)	Not specified	Street address, service dates, birth dates, ages > 89, death dates	N/A	Only state govt. agency and select entities working with the state	N/A
Tennessee Division of Health Planning	Tennessee All-Payer Claims Database (http://health.tn.gov/HealthPlanning/index.shtml)	Enrollment information; medical and pharmacy claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicaid (managed care, no FFS Medicaid in state) • There are plans to include data from the Medicare program in the future	Not specified	5-digit ZIP code, service dates, birth dates (year only), ages > 89	N/A	State govt. agency	N/A

Notes: (CMS) Centers for Medicare & Medicaid Services; (DUA) data use agreement; (FFS) fee-for-service.
Data are based on information available as of November 2014.
[a]States in most cases do not use the research identifiable file and limited data set terminology. The current listings come from the language found on state websites.

Table 3. Characteristics and Restrictions on Access to Selected State Funded All Payer Claims Data Sets (APCDs) (Cont'd)

AGENCY NAME	DATA SET NAME AND WEBSITE	BRIEF DESCRIPTION OF DATA	TYPE OF FILE[a]	PROTECTED HEALTH INFORMATION	RESTRICTIONS ON DATA ACCESS		
					PURPOSE OF DATA REQUEST	TYPE OF REQUESTOR	TYPE OF FUNDING SOURCE
Colorado Dept. of Health Care Policy and Financing (HCPF) and Center for Improving Value in Health Care	Colorado All-Payer Claims Database (http://www.civhc.org/All-Payer-Claims-Database/Data-Release-Review-Committee.aspx/)	Enrollment information; medical and pharmacy claims; and provider information from the following: • Commercial payers • There are plans to include self-funded and third party administrators • Medicare (Medicare Advantage, Medicare Part D) • Medicaid (managed care, almost no FFS in state) • There are plans to include FFS data from the Medicare program in the future.	Limited Data Set	5-digit ZIP code, service dates, birth dates, ages > 89	Only research supporting the Colorado Triple Aim of better health, better care, and lower costs.	No explicit restrictions	No explicit restrictions
			Identifiable Information Data Set	Street address, service dates, birth dates, ages > 89	Comparative effectiveness research	Researchers and health care providers	No explicit restrictions
Maryland Health Care Commission (MHCC)	Maryland Medical Care Database (http://mhcc.dhmh.maryland.gov/irb/Pages/default.aspx)	Enrollment information; medical, pharmacy, and dental claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicare (FFS under DUA with CMS, Medicare Advantage, Medicare Part D) • Medicaid	Not specified	5-digit ZIP code, service dates, birth dates (year and month only), ages > 89	Purpose of Medicare FFS data must fall under the DUA with CMS. No explicit restrictions for other data	Only the MHCC can use Medicare FFS data. No explicit restrictions for other data	No explicit restrictions
Office for Oregon Health Policy and Research	Oregon All Payer All Claims Database Limited Data Set (http://www.oregon.gov/oha/OHPR/RSCH/pages/apac.aspx)	Enrollment information; medical and pharmacy claims; and provider information from the following: • Commercial payers • Self-funded and third party administrators • Medicare (Medicare Advantage, Medicare Part D) • Medicaid (FFS and managed care)	Limited Data Set	5-digit ZIP code, service dates, birth dates (year only), ages > 89	Only if purpose serves the public interest and supports the mission and aims of the Oregon Health Authority	No explicit restrictions	No explicit restrictions

Notes: (CMS) Centers for Medicare & Medicaid Services; (DUA) data use agreement; (FFS) fee-for-service.
Data are based on information available as of November 2014.
[a]States in most cases do not use the research identifiable file and limited data set terminology. The current listings come from the language found on state websites.

Volume 4 (2016) Issue Number 2

Data costs also vary widely across state APCDs (Table 4). New Hampshire and Maryland do not charge a fee, whereas Oregon charges a fixed fee per file type per year of data to all requestors. Three APCDs (Colorado, Massachusetts, Maine) that permit access to external researchers use tiered pricing with higher fees for commercial entities and lower fees for academic and nonprofit institutions. Fees paid by nonprofits and academic institutions are typically one fourth to one half of the fees paid by other entities. Colorado is unique in that it has appropriated $500,000 to offset costs related to purchasing the APCD for academic institutions, state agencies, and nonprofits with revenue less than $5 million.

In order to request APCD data, Colorado, Massachusetts, Maryland, and Oregon require a project summary, data management plan, DUA, and variable selection sheet; however, procedures vary considerably. Oregon, Maryland, and Colorado (for its more PHI-sensitive file version) also require IRB or privacy review board approval. In terms of the recency of data available, APCDs from Maine, Maryland, and New Hampshire include close to real-time data, whereas the remainder lag by three to four years.

Discussion

Our review is the first to examine access and request policies for a broad cross section of publicly funded data sets available to researchers in the United States for analyses in areas such as comparative effectiveness research; patient-centered outcomes research; and access, cost, and quality of care research. We reached three important conclusions. First, there is significant heterogeneity in data access policies across the databases we reviewed, at both the federal and state levels. Second, access restrictions are not consistently related to the type or level of individual or protected health information (PHI) available in the data. Finally, even when data access is permitted to external researchers, there are various indirect barriers to access in the form of complex application procedures, fees, and data management requirements. We discuss each of these issues in further detail below.

We found significant variation in data access policies across federal agencies as well as for APCDs across state health agencies. In some cases, there was even variation in access policies across multiple data sets available from the same agency. This was particularly the case with CMS, wherein the variation stemmed from their designation of files based on the level of individual health information included. For example, despite the similar Medicare Part A and B claims data included in their Chronic Conditions Data Warehouse database and Standard Analytic Files, the one with geographic detail at the five-digit ZIP code level is designated as an RIF, whereas the one with county codes is classified as an LDS. In other cases, the LDS-RIF distinction is blurred. CMS's Medicare Current Beneficiary Survey contains the same PHI data elements as CMS's CCW files, but the former is designated an LDS and the latter is classified as an RIF. Although all the RIFs maintained by CMS should be classified as LDS files—based on the definition of LDS found in the HIPAA Privacy Rule—these distinctions in file types used by CMS are important, as they can have a major impact on the availability of individual-level data needed for high quality research, on whether or not certain types of researchers (particularly those funded by commercial entities) have access to the files, and also on the stringency of the data request approval process and data security requirements. In addition, policies from CMS have an impact on a broader array of data sets (e.g., the NIH's Health and Retirement Study or certain state APCDs) when those data sets are linked to Medicare data. They may also have an impact on new data infrastructure initiatives focused

Table 4. Data Cost, Request Process and Timelines, Data Storage, and Recency for State All Payer Claims Data Sets (APCDs) Available to External Users

NAME OF DATA SET	TYPE OF FILE	DATA COST[a]	DOCUMENTS AND APPROVALS REQUIRED TO SUCCESSFULLY REQUEST THE DATA SET						TIMELINESS OF DATA REQUEST APPROVAL	DATA STORAGE AND ACCESS POLICIES	MOST RECENT YEAR OF DATA AVAILABLE
			PROJECT SUMMARY	DMP	DUA	VARIABLE SELECTION WORK-SHEET	IRB APPROVAL	PRIVACY BOARD REVIEW			
Massachusetts All-Payer Claims Database	Level 2 File	Fixed fee per file type of data depending on requestor: academic, $$$; and others, $$$	Yes	Yes	Yes	Yes	No	No	~12 to 20 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	2012
Maine All-Payer Claims Database		Fixed fee per file type per year of data depending on requestor: commercial, $$$; assessed (e.g., provider, health insurer, etc.), $$$; nonprofit and educational, $$; redistributor, $$$	No	No	Yes	No	No	No	At least 6 weeks for nonpractitioner-identifiable data; at least 8 weeks for practitioner-identifiable data	Physical data files mailed to researcher (no specific policies on securing research environment)	Commercial & Medicaid (2014) Medicare FFS (2013)
New Hampshire Comprehensive Health Care Information System	Commercial Limited Use Data Set	No cost	Yes	No	Yes	Yes	No	No	~12 weeks	Physical data files mailed to researcher (no specific policies on securing research environment)	2014
Colorado All-Payer Claims Database	Limited Data Set	Licensing fee based: nonprofit with revenues $3–$5 mil, $$$; nonprofit with revenues of $1–$3 mil, $$$; nonprofit with revenues <$1 mil, $$$; academic institutions, $$$; state agencies,$$; other organizations, $$$[b]	Yes	Yes	Yes	Yes	No	No	~4 to 11 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	2012
	Identifiable Information Data Set		Yes	Yes	Yes	Yes	IRB approval, Privacy Board Review Approval, or proof of patient authorization			Physical data files mailed to researcher (rigorous policies on securing research environment)	

Notes: (DMP) data management plan; (DUA) data use agreement; (FFS) fee-for-service; (IRB) institutional review board.
Data are based on information available as of November 2014.
[a]$ denotes typically < $1000; $$ denotes typically >=$1000 and < $10,000; $$$ denotes typically > $10,000.
[b]Academic institutions, state agencies, and nonprofits with revenue less than $5 million are eligible for external APCD scholarship support that covers almost two-thirds of the cost, on average.

Table 4. Data Cost, Request Process and Timelines, Data Storage, and Recency for State All Payer Claims Data Sets (APCDs) Available to External Users (Cont'd)

NAME OF DATA SET	TYPE OF FILE	DATA COSTA	DOCUMENTS AND APPROVALS REQUIRED TO SUCCESSFULLY REQUEST THE DATA SET						TIMELINESS OF DATA REQUEST APPROVAL	DATA STORAGE AND ACCESS POLICIES	MOST RECENT YEAR OF DATA AVAILABLE
			PROJECT SUMMARY	DMP	DUA	VARIABLE SELECTION WORK-SHEET	IRB APPROVAL	PRIVACY BOARD REVIEW			
Maryland Medical Care Database		No cost	Yes	Yes	Yes	Yes	Yes	No	At least 4 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	2014
Oregon All Payer All Claims Database Limited Data Set	Limited Data Set	Fixed fee per file type per year of data ($$)	Yes	Yes	Yes	Yes	Yes	No	At least 6 weeks	Physical data files mailed to researcher (rigorous policies on securing research environment)	2011

Notes: (DMP) data management plan; (DUA) data use agreement; (FFS) fee-for-service; (IRB) institutional review board.
Data are based on information available as of November 2014.
[a]$ denotes typically < $1000; $$ denotes typically >=$1000 and < $10,000; $$$ denotes typically > $10,000.
[b]Academic institutions, state agencies, and nonprofits with revenue less than $5 million are eligible for external APCD scholarship support that covers almost two-thirds of the cost, on average.

on patient-centered outcomes research (PCORnet) and postmarketing surveillance (IMEDS Research Lab).

CMS recently relaxed access restrictions to RIFs for commercial entities by allowing innovators and entrepreneurs to request and use Medicare and Medicaid RIFs through the Virtual Research Data Center (VRDC).[8] The purpose of the data request may be related to business operations or research that leads to the creation of products or tools that the requestor intends to sell. In addition to the standard process for approval of a research protocol under the existing research request process, innovator and entrepreneur requests require an additional level of review that will examine whether the use of the data could hurt beneficiaries or lead to fraud and abuse in CMS programs. This innovator and entrepreneur access policy is a significant shift in CMS's philosophy toward data access for commercial entities, and it will be interesting to observe how the policy evolves in the near future as requests are submitted and reviewed.

Some of the inconsistency in data access policies across agencies may stem from the fact that several federal laws have implications for privacy and data access; there is no single, coherent framework to govern data use decisions.[9] Some efforts to simplify this situation are underway, including proposed revisions to the federal Common Rule that seek to clarify and harmonize regulatory requirements and agency guidelines.[10] State and federal laws may also overlap or come into conflict, in which case the one with stricter privacy protections typically takes precedence. At the state level, differences in access policies may also reflect the range of intended uses of the data (e.g., for health department reports in Minnesota versus "to improve the health of Maine citizens" in Maine).[11] Other groups have released

reports offering guidance to states seeking to create or revise APCDs; this may lead to greater consistency in state APCD access policies.[12]

We also observed numerous indirect hurdles to obtaining access to publicly funded databases that hinder use by all researchers. For instance, combining complex application procedures and high data user fees for the CCW (Medicare) and MAX (Medicaid) databases constrains access given the increasingly tight funding environment. Furthermore, the time from data request to approval and final data delivery may interfere with the ability to conduct and publish timely research. While LDS files available from CMS are cheaper with fewer complex applications and more timely delivery, these also raise issues for researchers. Specifically, some LDS do not contain complete utilization claims in one data set (e.g., Medicare SAF files exclude Medicare Part D data), thereby severely limiting the usefulness of these files. Others (e.g., the MCBS) contain full claims records but have insufficient sample size for health outcomes research. The differences in cost and time lag may be associated with the costs and resources required to generate minimum necessary data for specific research, the inability for some organizations to be remunerated for data services, or lack of workforce to effectively de-identify information using statistical methods. Given the rapid pace of medical innovation and quality improvement strategies in recent years, old data quickly become irrelevant and barriers that deter applicants or delay analysis have the potential to significantly curtail data utility and delay publication of important research findings.

Study Limitations

Four study limitations warrant mention. First, while our selection of data sets was not exhaustive, we believe it to be representative of the larger universe of publicly funded United States health data sets that contain individual-level PHI variables and sufficient detail for conducting high quality research studies. Second, the information presented here was current as of November 2014, but readers should be forewarned that rules governing data accessibility are dynamic and subject to change. Third, although the accuracy and quality of secondary databases is critical to their utility in health services research, evaluating data integrity was outside the scope of our review. Fourth, while these policies are largely representative of data access associated with claims data, future policies for clinically rich data from electronic health records and mobile health are likely to refer to existing policies for precedence. Limited data and hurdles to the exchange of information are likely to dampen the promise and ability of electronic medical records and mobile health data to be shared across sites, providers, and systems to enable and incentivize high-quality care and reduce duplicative services.

Recommendations and Conclusions

Our analysis highlights the need for policymakers and data owners to carefully evaluate their data access and request policies and to remove unnecessary barriers to the utilization of these valuable resources. Several revisions to current policies may be useful. First, placing a greater emphasis on research quality and intent, rather than simply the investigator's affiliation, may create greater opportunities while maintaining the spirit of patient protection. There is little disagreement that certain commercial uses of these publicly funded data sets would be inappropriate (such as to aid marketing of a product or service), yet the notion of "commercial interest" has changed over time, given that provider groups, hospitals, and health plans could potentially use data to conduct research on factors that improve outcomes and performance, which can in turn influence both quality of care and profits. This thinking is reflected

in a recent CMS policy revision, discussed earlier, which expands access to RIFs to the private sector.[9,14] Shifting the focus to a user's ability to submit a research proposal that is of high quality (i.e., methodologically rigorous), with scientific justification for including specific data elements (e.g., more detailed geographic data that are needed to investigate a specific hypothesis), and that offers promising utility (e.g., the potential to improve program administration or the health of the covered population), may be more important than blanket inclusion or exclusion of researchers affiliated with an academic institution or commercial entity. For example, a pharmaceutical company may wish to use data for research designed to identify modifiable factors influencing medication adherence in order to strengthen patient support resources, or a hospital might wish to conduct research on high-risk patients and to discern key factors that reduce readmission rates. As with other areas of research, consistent policies (such as the International Committee of Medical Journal Editors guidelines that govern authorship in medical journals) and transparency (such as the requirement to post study information on clinicaltrials.gov) can increase public accountability, promote research rigor, and increase the generalizability of knowledge while addressing potential conflicts of interest. Data use policies might similarly include requirements to agree to agency stipulations regarding use and handling of data, to publicly post study protocols, and to publish or release all results within a specified period. Perceived financial conflict of interest could be addressed by permitting greater access through third parties that are not necessarily academic or nonprofit, as long as the involved parties meet certain agency criteria for independence.

Second, increasing the availability of research-identifiable files may aid efforts to accelerate our understanding of precision medicine and to evaluate whether delivery system reform is having the intended impact on health care quality, effectiveness, and value. While LDS files may be sufficient for many questions, the inability to link to other files will leave many research questions unanswered. As information shifts from claims to clinically rich electronic health records, the ability to link across data sets will be increasingly important.

Third, remote access may ease the concerns of other stakeholders concerned with broader access to individual identifiable information. For example, resources such as the IMEDS Data Lab and CMS Virtual Research Data Center enable remote access to RIFs in a secure environment, reduce the need for one-off storage by individual researchers, and allow individual-level analysis while limiting the ability to download study results to the aggregate (rather than individual) level.

Finally, alleviating direct access barriers may reduce indirect barriers as well. For example, a tiered pricing approach for data user fees—whereby commercial entities with greater resources are charged higher fees that help offset lower fees collected from academic and nonprofit organizations—could enable more investigators to explore important research questions. Expanding access could also have the potential to reduce time lag barriers by generating revenue to support increased staffing and more timely processing. This is the approach taken by the United Kingdom's Clinical Practice Research Datalink (CPRD) GOLD, a publicly funded, primary-care research database of computerized medical records from across the United Kingdom.[15]

Greater consistency and simplicity in agency policies have the potential to increase access to, and thereby the utility of, a wide spectrum of publicly funded health care data sets. Further, greater consistency regarding the treatment of PHI, including reconciliation of federal laws, is needed. Finally, a framework that focuses more on user qualifications and intent across stakeholder groups, as opposed to simply user affiliation, would enable research that evaluates our current progress and moves us closer to our goals for a health care system that rewards quality, efficiency, effectiveness, and value.

Acknowledgements

This study was funded by the National Pharmaceutical Council.

References

1. Fung V, Brand R, Newhouse J, Hsu K. Using Medicare data for comparative effectiveness research--opportunities and challenges. Am J Manag Care. 2011 Jul;17(7):488-96.
2. The White House. Open Government Initiative [Internet]. Washington (DC): The White House; [updated 2015 April 14; cited 2015 April 14]. Available from: http://www.whitehouse.gov/open.
3. APCD council. Interactive State Report Map [Internet]. Durham (NC): APCD council; c2009-2015 [updated 2015 April 14; cited 2015 April 14]. Available from: http://apcdcouncil.org/state/map.
4. The National Patient-Centered Clinical Research Network. About PCORnet [Internet]. Washington (DC): The National Patient-Centered Clinical Research Network; [updated 2015 December 21; cited 2015 December 23]. Available from: http://www.pcornet.org/about-pcornet/.
5. Reagan-Udall Foundation for the Food and Drug Administration. Innovation in Medical Evidence Development and Surveillance (IMEDS) [Internet]. Washington (DC): Reagan-Udall; c2013 [updated 2015 December 23; cited 2015 December 23]. Available from: http://imeds.reaganudall.org/AboutIMEDS.
6. Centers for Medicare and Medicaid Services. Medicare provider utilization and payment data: Part D prescriber [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [updated 2015 April 30; cited 2015 June 1]. Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html.
7. 45 Code of Federal Regulations. Parts 160, 164 as amended by the Health Information Technology for Economic and Clinical Health (HITECH ACT).
8. Other requirements relating to uses and disclosures of protected health information, 45 C.F.R. Subtitle A § 164.514 (2011).
9. Centers for Medicare and Medicaid Services. Innovator Research [Internet]. Baltimore (MD): Research Data Assistance Center; [updated 2015 September 14; cited 2015 September 14]. Available from: http://www.resdac.org/cms-data/request/innovator-research.
10. Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. J Comp Eff Res. 2015;4(1):61-74.
11. Department of Health and Human Services (US). Human subjects research protections: enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators. Proposed rules. Fed Regist. 2011 Jul 26;76(143): 44512-31.
12. Freedman Healthcare. All payer claims database workgroup: recommendations to the Minnesota legislature [report on the Internet]. Minneapolis (MN): Minnesota Department of Health; 2014 December 24 [cited 2015 Apr 14]. Available from: http://www.health.state.mn.us/healthreform/allpayer/APCDwkgrpFinalRpt2015Jan.pdf.
13. Porter J, Love D, Costello A, Peters A, Rudolph B. All-Payer Claims Database Development Manual: Establishing a Foundation for Health Care Transparency and Informed Decision Making [report on the Internet]. Durham (NH): APCD council; 2015 Mar [cited 2015 April 14]. Available from: http://apcdcouncil.org/sites/apcdcouncil.org/files/All-Payer%20Claims%20Database%20Development%20Manual_03042015_0.pdf. Joint publication with the West Health Policy Center.
14. National Institute for Health Research (UK). Accessing CPRD data [Internet]. London, England (UK); National Health Service; [cited 2015 Apr 14]. Available from: http://www.cprd.com/dataAccess/default.asp#OnlineDataGOLD.