# Overcoming Barriers to a Research-Ready National Commercial Claims Database

David Newman, JD, PhD; Carolina-Nicole Herrera, MA; and Stephen T. Parente, PhD


## ABSTRACT

### Objectives

Billions of dollars have been spent on the goal of making healthcare data available to clinicians and researchers in the hopes of improving healthcare and lowering costs. However, the problems of data governance, distribution, and accessibility remain challenges for the healthcare system to overcome.

### Study Design

In this study, we discuss some of the issues around holding, reporting, and distributing data, including the newest "big data" challenge: making the data accessible to researchers and policy makers.

### Methods

This article presents a case study in "big healthcare data" involving the Health Care Cost Institute (HCCI). HCCI is a nonprofit, nonpartisan, independent research institute that serves as a voluntary repository of national commercial healthcare claims data.

### Results

Governance of large healthcare databases is complicated by the data-holding model and further complicated by issues related to distribution to research teams. For multi-payer healthcare claims databases, the 2 most common models of data holding (mandatory and voluntary) have different data security requirements. Furthermore, data transport and accessibility may require technological investment.

### Conclusions

HCCI's efforts offer insights from which other data managers and healthcare leaders may benefit when contemplating a data collaborative.


*Am J Manag Care. 2014;(11 Spec No. 17):eSP25-eSP30*

Big healthcare data have become ubiquitous, and discussions about such data often focus on outcome metrics and healthcare costs. However, the jump from a raw database to an analytical file is fraught with issues. There are problems of governance, data distribution, and accessibility that data holders must overcome before researchers and policy makers can benefit from the data.



In this study, we discuss how one research-oriented, data-holding organization, the Health Care Cost Institute (HCCI), has addressed some of the barriers to effective data use.[1] HCCI is a nonprofit, nonpartisan, independent research institute that serves as the repository of healthcare claims for more than 50 million Americans per year (for 2007 through 2013) from 3 of the nation's largest insurers. HCCI holds individual insurance, group insurance, and MedicareAdvantage data in a manner compliant with the Health Insurance Portability and Accountability Act (HIPAA) and antitrust law, and in a manner that addresses insurers' concerns about company confidentiality. HCCI licenses and distributes data to research institutions. HCCI's current challenge is building a scalable, secure data distribution system to support timely, independent healthcare research. Below, we describe HCCI's approach to data governance and distribution, and we explore ways in which technology may help data holders promote public research.

## BACKGROUND

As recently as 6 years ago, data on healthcare costs were relatively scarce. Some states, through either local initiatives or national efforts, had hospital reporting on healthcare utilization.[2] Other states had launched efforts to mandate the reporting of healthcare claims from private insurers.[3] Some organizations, such as Blue Cross Blue Shield or Thompson Reuters (now Truven Health Informatics), had commercialized private healthcare data from a limited set of insurers and/or employers.[4]

**Take-Away Points**

Research-focused data organizations face challenges in holding and distributing healthcare data. Using the experience of the Health Care Cost Institute as a model, the authors explore approaches and obstacles to making claims available to researchers.

- There are 2 common models for holding multi-payer health claims data.
- Date security requirements will vary depending on the model.
- Data licensing requirements will be complicated by the Health Insurance Portability and Accountability Act, anti-trust regulation, and other issues.
- Two external issues, data transport and access, may require more technological investment to make data research-ready.

Today, healthcare data, in general, are more available. The federal government has a number of ongoing initiatives. HHS began an effort to build a national multi-payer claims database to support comparative effectiveness research.[5] CMS embarked on efforts to make Medicare data more available to the states and launched the Qualified Entity program.[6] CMS also invested in the creation of a virtual research data center to make Medicare data more accessible to researchers. Additionally, the Affordable Care Act and the American Relief and Recovery Act increased provider use of electronic health records.[7,8] Some states mandated all-payer claims databases (APCDs) to support insurance regulation and inform public health policies.[4] Other states, in particular Hawaii and Arkansas, have used federal dollars available from the Center for Consumer Information and Insurance Oversight (CCIIO) to initiate public reporting efforts based on claims data.[9] Employers and communities are also collecting and sharing data on local healthcare markets.[10,11] The Midwest Health Initiative holds health data on some 1.8 million residents of St. Louis, Missouri, and 18 nearby counties.[6] The California Healthcare Performance Information System (CHPIS) collects data and reports on physicians.[7]

There is more reporting of healthcare prices to consumers through the Internet. FAIR Health, a nonprofit, "offers unbiased data products and services to consumers, the healthcare community, employers, unions, government agencies, policymakers, and researchers."[12] Castlight Health, Inc, offers transparency services to the employees of businesses enrolled in their system "to enable employers and health plans to lower the cost of healthcare and provide individuals unbiased pricing and quality information to make smart healthcare purchase decisions."[13,14] Though both these initiatives hold multi-state data from multiple data suppliers, their respective scopes limit access to their pricing information. When data are available for the patient's location, FAIR Health provides healthcare cost information to the public via billing codes. Castlight provides their clients' employees with price transparency services but does not provide this information to the public.

Perhaps unique among these efforts is HCCI, a nonprofit research institute. HCCI was launched in 2011 with the support of 4 of the nation's largest national health insurance companies: Aetna, Humana, UnitedHealthcare, and Kaiser Permanente.[1] Overseen by an independent governing board principally composed of academic economists, the institute's public mission is to report on trends in cost and utilization, to make commerical claims data available for research and, most recently, to help states.[1] To support these missions, HCCI assembled a national multi-payer claims database with allowed amounts (actual prices paid to providers for services). HCCI has released several reports describing national trends in healthcare spending, prices, and utilization, and has provided statistically de-identified databases to research institutions for noncommercial purposes. To do so, HCCI has addressed 2 of the key barriers to using big healthcare data: governance and distribution.

## GOVERNANCE

Health data organizations, whether they are public or private entities, face challenges in holding healthcare data. Foremost is getting permission from the owners of data (healthcare payers, providers, and patients) to assemble a database (data contribution). Then, organizations face a series of challenges related to keeping the information private and useful. The way an organization collects data changes the way it responds to these challenges.[14]

### Mandatory Contribution Models

Most holders of multi-payer health data receive their data through a mandatory contribution model (MCM). MCMs occur when a state government requires that insurers, providers, and/or employers provide healthcare data for statutory purposes (such as insurance or provider regulation).[4] A common form of MCM is the mandated state all-payer claims database. Many MCMs require data owners to provide healthcare claims using unique data extraction rules, such that many states operate with different data specifications.[15] Not surprisingly, this sort of effort is costly. States face significant costs as they develop customized solutions and analytic results.[16] For example, for fiscal year 2015, Maine's MCM-governed ACPD may cost the state about $1.66 per Maine resident.[17] Moreover,

multiple data feeds and multiple reporting systems burden providers and payers who cross state boundaries. How much MCM compliance costs providers and payers has not been documented.

### Voluntary Contribution Models

An alternative to the MCM is the voluntary contribution model (VCM). VCMs are a contractual approach in which data owners voluntarily contribute information to a data collaborative. Many VCMs are associated with not-for-profit entities such as HCCI, the Wisconsin Health Information Organization, and the Midwest Health Initiative.[6,18] A growing number of states are considering a VCM, and in at least 1 (Virginia), the Commissioner of Insurance has negotiated insurer participation in a statewide data-sharing effort.[19] VCMs require greater confidence-building than MCMs, as the entities providing the data voluntarily relinquish some control of their data. For example, HCCI and its data contributors entered into a series of agreements that govern how the research institute can use these data and the terms under which it can license these data. HCCI maintains an internal data integrity committee whose mandate is to ensure that HCCI's activities conform to the law and to its contractual obligations concerning the data. Even with the cost of coordination, this type of effort could be less costly than an MCM. A national or multi-state VCM is very scalable for both the VCM and for data owners as long as each contributor sends 1 feed for all geographies with a common set of data definitions and requirements. Scalability for a VCM declines if it is restricted to the state or sub-state level, or if the VCM has not sufficiently invested in data standards. However, VCMs may not have the utility for some state purposes as MCMs do, in part because consensus, not statute, dictates data use.

### Transitional Contribution Models

A number of states are also contemplating developing hybrid contribution models wherein data contribution would begin as a VCM and then transition to an MCM. These transitional contribution models (TCMs) would operate initially as VCMs and, after a start-up period, transition into MCMs. Helping drive the emergence of the TCMs are CCIIO grants to support rate reviews and price transparency.[20] Both Arkansas and Hawaii have indicated they are interested in developing a TCM using the Cycle III grant funds. Arkansas has released a request-for-proposal to support building its data center as a VCM and then transitioning it to an MCM.[9] In December 2013, Hawaii asked commenters to detail potential issues involved with moving from a VCM to an MCM, including questions regarding sustainability, data owner relationships, and data standards.

One key governance issue with TCMs is data licensing. When contemplating a TCM, one should be aware that data licenses are not necessarily transferable if the data holding organization's legal structure changes. For example, if the model begins as a nonprofit effort (like HCCI) and is then integrated with a state or federal agency, the data licenses will likely need to be renegotiated, and currently held data may have to be destroyed. If a VCM becomes an MCM without changing legal structure, a data holder may avoid this potential legal hazard.

### Data Privacy and Confidentiality

Depending on the data contribution model, the data holder may have different protected health information requirements. To ensure privacy, VCMs often face more restrictions than do MCMs. Qualified entities and state agencies are not as restricted by HIPAA as other data holders.

HIPAA provides federal protections for personal health information and constrains the data available for research.[2] However, HIPAA generally provides that health information is not individually identifiable if someone, with appropriate training and accepted statistical and scientific methods, determines that the risks of identifying an individual in the data are small. Using this framework, HCCI uses and distributes "statistically de-identified databases." HCCI currently distributes 2 statistically de-identified claims "data views," distinguished by the protected health information allowed in each. For example, one view has year of birth, whereas the other has patient zip code. To maintain statistical de-identification, research teams may not combine or merge the data views. This approach allows researchers to receive the richest claims data possible in an HIPAA-compliant manner.

In addition to HIPAA, everything that an MCM, a VCM, or a TCM does must conform to applicable antitrust law. In the case of HCCI, there is a 1-way flow of the data from the data contributors to HCCI. The data contributors have no rights to the combined database or any access to the combined database. Every research product generated by HCCI undergoes a legal review for antitrust issues. Finally, HCCI does not perform any proprietary or confidential research using its data on behalf of the data contributors.

## Data Use

Once governance around data holding has been satisfied, data holders need to address the issues of data use. License agreements, at a minimum, need to deal with the HIPAA. In addition, data holders often address the rights to publish and ownership of any intellectual property developed. Finally, the inadvertent or intentional release of company confidential information is of particular concern to data contributors, regardless of contribution model, and may require scrutiny of research and data products.

For most VCMs and MCMs, data are licensed after a proposal is made by researchers. Typically, a research committee reviews the proposal and may require research to be governed by an institutional review board. At HCCI, a scientific review committee reviews all research proposals, including those with funding from peer-reviewed institutions such as the National Institutes of Health. This committee is composed solely of academic researchers, and data contributors have no representation on the committee. HCCI data use, furthermore, is limited to the proposed purposes.

Unlike most VCMs or MCMs, HCCI does not license data directly to the researcher. Rather, HCCI licenses data to the researcher's university on behalf of researchers. Licensing data to universities can make the data more widely available to research teams. HCCI's Academic Research Partnerships allow a university to license multiple projects (including student projects) per year. In addition to saving the time of negotiating individuals' data licenses, this allows a university seeking to support or build a research program around healthcare claims data to do so more efficiently. As discussed elsewhere here, it would also allow universities that do not have sufficiently robust research technologies to leverage the technologies HCCI has developed for data distribution and research project management. The result is secure uses of the data by multiple research teams.

However, licensing data to universities is not as straightforward as one might think, as university attorneys are inclined to want to renegotiate standard license terms. Because HCCI is a VCM, certain terms are simply not negotiable, and these constraints must be recognized in the license. Some clauses, such as choice of jurisdiction, are quickly resolved. Issues that tend to slow down the licensing process are confidentiality provisions and intellectual property rights.

To protect against the intentional release of confidential information, MCMs and VCMs have to take steps to reduce potential violations. HCCI developed a set of masking rules generally designed to deal with how prices are publicly reported, as these are the salient pieces of information that give rise to antitrust concerns. These rules do not constrain research but do prohibit reporting of analyses at the data contributor level. As is typical with health data, the rules allow raw reporting of specific service prices within a specific geography when the data meet a critical threshold of observations. When researchers who want to report on a specific service in a specific geography do not meet the thresholds, they are required to either expand the geographic area, select a different geographic area to highlight, or aggregate the service data.

## Distribution and Accessibility

After governance issues, the data holder faces 2 major technical challenges to distributing data in a world of relatively rich storage options: transport and updating. Data holders whose purpose is to help inform knowledge about healthcare also face another challenge—making data accessible to research teams who lack the current means of processing large data. As HCCI's public mission is to promote research and reporting of healthcare costs and utilization, it has had to address these challenges and is developing innovative solutions.

## Transport

After the data are licensed and are ready for distribution, the data holder needs to transport the data to the end user. Depending on the size of the data, the technical capacity of the data holder, and the technical capacity of the data recipient, 2 forms of transport are commonly used: physical transport of a data drive through a courier service or electronic transmission through a secure gateway.

Physical drive transmission is very common with Medicare data, and Buccaneer, the Medicare vendor, physically transports secure files to research teams around the country.[21] The Agency for Healthcare Research and Quality also provides physical copies of hospital inpatient data.[2]

Less common is the use of electronic transfer gateways, although this method is gaining popularity, particularly as costs decline. Gateways offer greater control over data transmission, require a direct link between repository and recipient, and require recipients to provide more detail about their data security. However, there are technological constraints to transmitting large databases over networks. For example, 1 year of employer-sponsored insurance claims data from HCCI are approximately 325 GB in a flat file. If the data are transmitted in a standardized analytic format (such as a *.sas7bdat or *.dta), the base file is much larger, which can make transmission on relatively weak connec-

tions impossible. In HCCI's experience, transmissions to major research universities work well in the range of less than 100 GB, resulting in multiple file segments that must be reassembled at the university. Transmissions to recipients who lack significant bandwidth can be prohibitively slow.

### Updating

The main benefit of a secure gateway is that it can also help data holders simplify the otherwise complex process of data updating. Unlike many other forms of data, health data are not static. Claims data go through 3 stages—filing, processing, and adjudication—with the timing of each dependent on the source of claims, the payer, and the patient. In the case of prescription claims, filing, processing, and adjudication can be accomplished on the same day. In the case of organ transplants, it may take more than 18 months to complete the adjudication.

For claims that are not fully processed, the data holder needs to decide whether it will offer raw, consolidated (detailed transactions with current claim payment statuses), or adjudicated (paid and final) claims.[22] Some APCDs update data holdings (using either raw or consolidated data) to include new payment information monthly or quarterly. Thus, a researcher who received data in January likely will have different data from another researcher who received data later in the year. One alternative is to provide only adjudicated claims data, which is the option that HCCI uses. As a result, filed but unpaid claims are not included. All data holders need a policy on run-out. If claims are collected and aggregated by year (be it calendar or fiscal), the data holder needs to decide how many months need to pass before it declares a year complete. In the case of HCCI, data are collected by calendar year with 6 months of run-out. This means that for care provided in 2007, HCCI does not receive data until 2008. HCCI also collects data with 12-, 18-, and 24-month run-out and, therefore, holds at least 99% of adjudicated claims.

Data holders will find the use and distribution of annual health data files complicated by changes in the data contributors. In an MCM, the number of data contributors should not retroactively increase as long as regulations do not change. In a VCM, the number of data contributors may change. HCCI has set as a policy that data should be retrospective from 2007 onward; therefore, a researcher requesting a data update will need file replacement if HCCI acquires more data contributors.

### Accessibility

Most academic research teams do not have the dedicated resources needed to store, process, and analyze large claims files. As of today, many claims data holders, including HCCI, cut customized files for researchers. Although some researchers may need only aggregated data, even highly aggregated databases can overwhelm the most advanced desktops. Successful users of "big health data" will need to invest in technology if other solutions are not available.

The researchers' challenges are also the data holders' challenges, particularly if the data holder is committed to supporting research. One solution is to push academic researchers to better leverage their existing infrastructures. As noted previously, HCCI is licensing data to universities and research institutions for use by multiple research teams. This partnership approach allows universities with data centers to use their processing assets for multiple projects over time and with a standard update schedule. Another solution is outsourcing the hosting for individual projects. HCCI may provide research teams with a set of suggested vendors whose security and technical requirements meet HCCI standards.

Alternatively, data holders who wish to promote research may need to invest in information technologies to make their data more widely available. This is the approach that HCCI is considering as a mechanism for advancing research and collaboration on healthcare data.[23] A robust virtual data research center could provide collaborative research teams with access to data within a secure environment. Such an environment would include 1) a query-ready database with limited data-merge capacity, 2) a secure portal by which authorized users can access the data, 3) secure and private storage for researchers, 4) isolated silos to keep research teams segregated and separated, 5) monitoring capacity, and 6) analytic tools. At this time, few vendors have both the processing power and analytic prowess to support research infrastructure.

## CONCLUSIONS

Advances in information technology have made it possible for healthcare researchers and other stakeholders to have access to greater healthcare data. Problems exist with the scale of the data, standards for data holding, governance, reporting and privacy, and rights to ownership. However, the greatest challenge—making "big data" accessible to research teams with great ideas but limited resources—remains unsolved. Future advances on the horizon may help to eliminate some of these concerns, but healthcare leaders should not expect an explosion of new data insights without investment in basic health services research infrastructure and technology.

***Author Affiliations:*** The Health Care Cost Institute (DN, C-NH), Washington, DC; Department of Finance, Carlson School of Management, University of Minnesota (STP), Minneapolis, MN.

***Source of Funding:*** None.

Author Disclosures: Dr Newman and Ms Herrera are employed by the Health Care Cost Institute. Dr Parente reports no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

***Authorship Information:*** Concept and design (DN, C-NH, STP); analysis and interpretation of data (C-NH); drafting of the manuscript (DN, C-NH); critical revision of the manuscript for important intellectual content (DN, C-NH); obtaining funding (STP); administrative, technical, or logistic support (DN, C-NH); supervision (DN, STP).

***Address correspondence to:*** Carolina-Nicole Herrera, MA, Director of Research, Health Care Cost Institute, Inc, 1310 G St NW, Suite 720, Washington, DC 20005. E-mail: cherrera@healthcostinstitute.org.

## REFERENCES

1. About HCCI. Health Care Cost Institute, Inc, website. http://www.healthcostinstitute.org/files/About%20HCCI%20-%209-19-12.pdf. Published 2013. Accessed May 16, 2013.

2. Overview of HCUP: Healthcare Cost and Utilization Project. HCUP/Agency for Healthcare Research and Quality website. www.hcup-us.ahrq.gov/overview.jsp. Published November 2009. Accessed May 16, 2013.

3. Miller PB, Love D, Sullivan E, Porter J, Costello A; for State Coverage Initiatives, Robert Wood Johnson Foundation. All-payer claims databases: an overview for policymakers. http://www.statecoverage.org/files/SCI_All_Payer_Claims_ReportREV.pdf. Published May 2010. Accessed March 3, 2014.

4. Hansen L, Chang S. Health Research Data for the Real World: The MarketScan Databases. Ann Arbor, MI: Thomson Reuters; 2012.

5. Navathe AS, Conway PH. Optimizing health information technology's role in enabling comparative effectiveness research. *Am J Manag Care*. 2010;16(12 suppl HIT):SP44-SP47.

6. Qualified Entity Program. CMS website. www.cms.hhs.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/QEMedicareData/index.html. Published October 2013. Accessed March 4, 2014.

7. Health reform implementation timeline. The Henry J. Kaiser Family Foundation website. http://kff.org/interactive/implementation-timeline/. Published 2013. Accessed September 19, 2013.

8. Blumenthal D. Launching HITECH. *N Engl J Med*. 2010; 362(5):382-385.

9. Awards and requests for proposals. All-Payer Claims Database Council website. http://apcdcouncil.org/awards-and-requests-proposals. Accessed March 3, 2014.

10. Our data. Midwest Health Initiative website. http://www.midwesthealthinitiative.org/our_data.php. Published 2013. Accessed May 15, 2013.

11. Multi-payer claims database (MCPD). California Healthcare Performance Information System website. http://www.chpis.org/programs/mpcd.aspx. Published 2014. Accessed December 15, 2014.

12. About us. FAIR Health, Inc website. http://www.fairhealth.org/About-FH. Published 2013. Accessed December 15, 2013.

13. Solutions. Castlight Health, Inc website. www.castlighthealth.com/solutions. Published 2013. Accessed May 16, 2013.

14. Interactive State Report Map. All-Payer Claims Database (APCD) Council website. http://apcdcouncil.org/state/map. Accessed December 15, 2014.

15. Costello A, Taylor M. Standardization of data collection in all-payer claims databases: fact sheet. APCD Council website. www.apcdcouncil.org/sites/apcdcouncil.org/files/Standardization%20Fact%20Sheet_FINAL_for010711release_1.pdf. Published January 2011. Accessed May 16, 2013.

16. Love D, Sullivan E. Cost and funding considerations for a statewide all-payer claims database: fact sheet. APCD Council website. www.apcdcouncil.org/sites/apcdcouncil.org/files/Cost%20Fact%20Sheet_FINAL_1.pdf. Published March 2011. Accessed May 16, 2013.

17. Human Services Research Institute (HSRI) Health Data Warehouse proposal prepared for the State of Maine, RFP #201207352. Maine Health Data Organization website. https://mhdo.maine.gov/_pdf/HSRI_HealthDataWarehouseProposal120824.pdf. Published August 27, 2012. Accessed March 3, 2014.

18. WHIO fact sheet 2013. Wisconsin Health Information Organization website. www.wisconsinhealthinfo.org. Published 2013. Accessed June 2013.

19. Health care prices. Virginia Health Information website. http://www.vhi.org/health_care_prices.asp. Accessed December 15, 2014.

20. The Center for Consumer Information & Insurance Oversight: grants to states to support health insurance rate review and increase transparency in health care pricing. CMS/CCIIO website. www.cms.gov/CCIIO/Resources/Fact-Sheets-and-FAQs/rr-foa-faq-5-8-2013.html. Published 2013. Accessed March 3, 2014.

21. Buccaneer for CMS. Chronic conditions data warehouse: Medicare administrative data user guide, version 2.0, 2013. www.ccwdata.org/cs/groups/public/documents/document/ccw_userguide.pdf. Accessed May 16, 2013.

22. Understanding the process (part 3 - implementation). OnpointCDM: Claims Data Manager website. www.onpointcdm.org/newsletters/newsletter_articles.php?id=5. Published 2009. Accessed March 3, 2014.

23. Newman D, Frost A, Herrera C, Parente S. The need for a smart approach to big health care data. HealthAffairs Blog website. http://healthaffairs.org/blog/2014/01/27/the-need-for-a-smart-approach-to-big-health-care-data/. Published January 27, 2014. Accessed March 3, 2014. ■

**www.ajmc.com** Published as a Web Exclusive