KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE

www.cdt.org

CENTER FOR DEMOCRACY
& TECHNOLOGY

1634 Eye Street, NW
Suite 1100
Washington, DC 20006

March 22, 2012

Decentralizing the Analysis of Health Data

As the digitization of health records makes it easier and more cost effective to share and analyze health data, policymakers and businesses are increasingly looking to use health data for secondary purposes – uses beyond that for which the health data were originally collected. For example, health data that were primarily collected for treatment or payment can be valuable for such secondary uses as population-scale research and public health surveillance. Done properly, many secondary uses of health data can provide substantial benefits to patients and aid the creation of a more effective, information-driven health care system.

Secondary use initiatives should be undertaken in a way that maximizes the confidentiality and security of patient data and preserves the trust of both health care providers and the public. While a strong policy framework based on Fair Information Practices is critical to achieve this balance, the technical architecture of information exchange – which is the focus of this paper – is another important factor.[1] Currently, many government programs using health claims for secondary purposes collect and retain the data in a centralized fashion. The key message of this paper is that decentralized alternatives can achieve most secondary use program goals in a manner that is more protective of privacy and security in the long term.

Whereas centralized databases typically operate by compiling data into one system and managing it from that location, decentralized systems typically leave data housed with the original sources of the data and perform analyses by searching the data held by these entities. Decentralized models are already being evaluated in some health care contexts.[2] One decentralized model, the "distributed query" system, may be adequate for many routine secondary use programs. Another approach, the "distributed access" system, is likely more appropriate for programs that may confer a competitive advantage or disadvantage on source of the data, such as where the program directly affects the revenue stream of the data source. Under the "distributed access" model, government agencies access structured data securely held by health care plans and providers, the agencies analyze the data to meet their compliance and research objectives, and the agencies keep the results of their analysis – but the agencies do not receive or retain a copy of the underlying data.

It should be noted that CDT is not urging federal or state agencies to immediately replace their existing centralized databases with distributed systems. Any decentralized system for

---

[1] Although this paper focuses on network architecture, that architecture – whether centralized or decentralized – will make little difference to patient privacy and data security if the policy and regulatory framework wrapped around the architecture is feeble or ineffective. See the Center for Democracy & Technology, *Comprehensive Privacy and Security: Critical for Health Information Technology*, May 14, 2008, pgs. 6-7, http://www.cdt.org/files/pdfs/20080514HPframe.pdf.

[2] For example, the Department of Health and Human Services has begun efforts to evaluate distributed systems through the Multi-Payer Claims Database (MPCD) project and the Food and Drug Association's Mini-Sentinel initiative.

population-scale data analysis faces technical challenges that must be overcome prior to deployment. CDT therefore recommends the U.S. Dept. of Health and Human Services (HHS) collaborate with health care plans and providers, researchers, state agencies, and technology vendors to initiate projects to evaluate the effectiveness of a distributed models.[3] Encouragingly, HHS recently released a final rule that establishes distributed architecture as the default for HHS' risk adjustment program.[4] Unfortunately, the regulations for some federal and state programs lock plans into a centralized model, which will make it more difficult to deploy decentralized approaches for secondary use programs in the future. CDT recommends federal and state agencies ensure their regulations leave open the possibility of using systems – subject to approval by the agencies – that do not rely on centralized databases.[5]

I. Existing secondary use programs tend to centralize data

Numerous state and federal government agencies have established programs that analyze health information to support data-driven health care reform and other policy objectives. At present, many of these programs utilize electronic health claims information held by payment plans, in part because those plans more routinely digitize claims information, whereas health data digitization among health care providers is uneven. However, analyzing clinical data from providers' electronic medical records (EMRs) is a longer-term goal for secondary use programs.[6] Analyzing clinical data – as opposed to merely claims – can support a wealth of valuable research programs, but will also make the associated data security and patient privacy issues more urgent. For purposes of this paper, we refer to data owners generically as "plans," though the network architecture of secondary use programs will ultimately affect providers as well.

Many claims analysis programs run on a "centralized" model whereby government agencies (or their contractors) collect individuals' health care claims from health plans, compiling the data

[3] CDT recommends HHS oversee projects that explore both distributed query and distributed access systems for population-scale health data analysis for secondary use programs. In the efforts underway so far, Mini-Sentinel is not directed at population-scale claims or medical record analysis. Mini-Sentinel operates on a distributed query model. The MPCD is still in an early planning stage, so it is not clear whether it will operate using queries or on an access model.

[4] The final rule does not go into detail regarding what that distributed architecture will look like.

[5] The Center for Budget and Policy Priorities recently came to a similar conclusion, recommending that HHS evaluate the distributed access approach to determine whether it can produce reliable analytics for purposes of the health reform law's risk adjustment program and to support other health care reform goals. Edwin Park, *Allowing Insurers to Withhold Data on Enrollees' Health Status Could Undermine Key Part of Health Reform*, Center on Budget and Policy Priorities, Dec. 12, 2011, pg. 6, http://www.cbpp.org/files/12-12-11health.pdf.

[6] See, e.g., Robert Rowley, *Anonymized EMR-based Data Analysis – The Next "Big Thing" in Healthcare*, Practice Fusion, EHR Bloggers, January 4, 2011, http://www.practicefusion.com/ehrbloggers/2011/01/anonymized-emr-based-data-analysis-next.html. See also, Netezza, Electronic Health Record Analytics & Secondary Use, http://www.netezza.com/data-warehouse-appliance-industries/health-record-analytics.aspx (last accessed February 15, 2012).

into one large-scale system and managing it from that location. In numerous cases, this centralized approach is required by regulation or legislation.

- Approximately 14 states have established centralized "all-payer claims databases" (APCDs) to compile digital claims data longitudinally for public policy, law enforcement, and research goals, including comparative effectiveness research.[7] Most – if not all – APCDs are required to operate on a centralized database model, either through program implementation requirements or state regulations.[8]

- The federal Office of Personnel Management (OPM) is in the process of building its "Health Claims Data Warehouse" for comparative effectiveness research, fraud detection, and other programs related to its management of the Federal Employee Health Benefits Program.[9] The regulations establishing OPM's Warehouse require the program to operate on a centralized database model.[10]

Recently, however, one major program has endorsed a decentralized approach. In July 2011, the Centers for Medicare and Medicaid Services (CMS) proposed a rule that would have compelled every state (or HHS on the state's behalf) to collect claims data from every payer in the individual and small group market to support the risk adjustment program required by the Patient Protection and Affordable Care Act of 2010.[11] The proposed rule would have required a centralized database approach for this program and would have denied states flexibility in adopting decentralized systems in the future.[12] In March 2012, in a very significant development, CMS issued its final rule and changed course from the proposed rule, setting distributed

---

[7] APCD Council, Interactive State Report Map, http://www.apcdcouncil.org/state/map (last accessed February 15, 2012).

[8] See, e.g., New Hampshire Revised Statutes Annotated, RSA 420-G:11(II)(a)(1)-(2), http://www.gencourt.state.nh.us/rsa/html/XXXVII/420-G/420-G-11-a.htm (last accessed March 12, 2012). See also New Hampshire Admin. Rules, Parts Ins 4003-4004, http://gencourt.state.nh.us/rules/state_agencies/ins4000.html (last accessed March 12, 2012).

[9] 76 Fed. Reg. 35050-4. Following the initial announcement of the Warehouse, CDT and other groups urged OPM to consider alternatives to centralization and to provide greater detail on the system's privacy protections. Center for Democracy & Technology, Letter to OPM Regarding the Health Claims Data Warehouse, October 27, 2010, http://cdt.org/files/pdfs/CDT_Letter_to_OPM_Re_Health_Claims_Data_Warehouse-102710.pdf.

[10] 76 Fed. Reg. 35050-4.

[11] 76 Fed. Reg. 41930, 41940-1. The proposed rule would have permitted state APCDs to perform this collection and analysis.

[12] *Id.* CDT issued comments to the proposed rule, arguing that CMS should adopt a distributed approach or – at minimum – not deny states the option of using a distributed approach in the future. Center for Democracy & Technology, Comments to CMS–9975–P, October 31, 2012, pgs. 5-6, http://cdt.org/files/pdfs/CDT_Comments_to_CMS-9975-P.pdf

systems as a default for the risk adjustment program.[13] The final rule provides states with the option to use their own approved models of claims collection and analysis for the program.[14]

II. The centralized model

For a more in-depth look at the structure and data flows of a secondary use program that utilizes a centralized database model, CDT examined APCDs, especially New Hampshire's Comprehensive Health Care Information System.

The authority of most APCDs to collect and release data stems from state laws and regulations requiring in-state entities – such as health plans – to submit data to the APCD or its designee, such as a data processor. In New Hampshire, state laws establishing the Comprehensive Health Care Information System require insurers to submit claims information to New Hampshire or its designee.[15] APCDs typically collect such data as health, pharmacy, and dental claims, as well as eligibility and provider files, from commercial and public payers. State APCDs usually collect data on all state residents. The health claims data include diagnosis codes, what procedures or treatments were billed, how the patient or plan paid for care, and the type of facility or provider submitting the claim.[16] Many states include Medicaid claims, and a growing number of states include Medicare claims as well.[17] APCDs often – but not always – want patient-specific information in order to compile longitudinal records on an individual patient basis. This allows the APCD to track individual patients as they change carriers and care settings over time. The inclusion of demographic information enables APCDs to break down patterns in cost, treatment, etc., by gender, location and more.

Currently all operational and developing state APCDs appear to have been built on a centralized data architecture model. Under this model, copies of the claims information are periodically transferred from health plans and providers into one large system controlled by the state, subject to the state's regulatory, policy, and technology frameworks. For the most part, APCDs accomplish this by establishing a secure file transfer protocol with health plans, through which the plans regularly submit claims data (typically monthly). Some state APCDs collect this data directly, though in other cases the plans submit the data to a third party designee before it is passed on to the APCD. For example, the New Hampshire Department of Health and Human

---

[13] 77 Fed. Reg. 17220, 17233.

[14] *Id.*

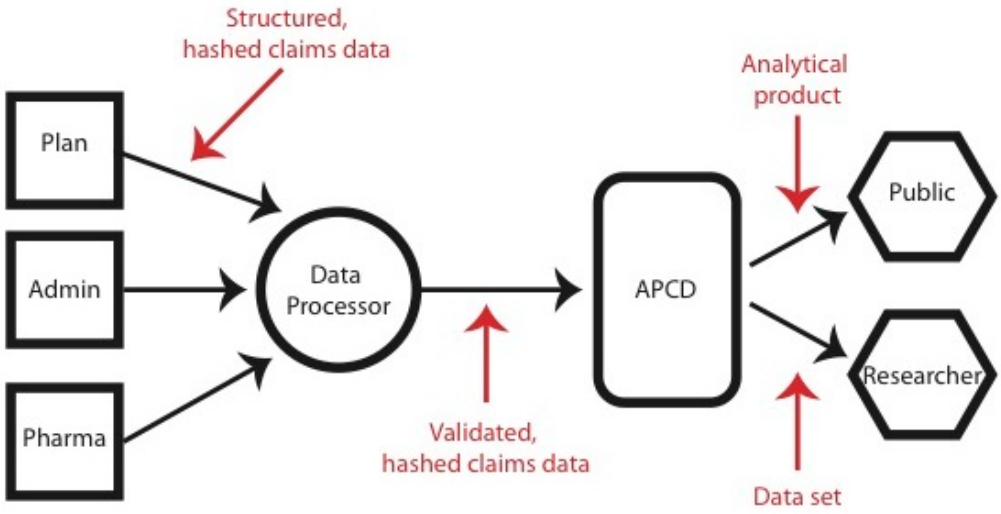
[15] New Hampshire RSA 420-G:11(II)(a)(1)-(2), http://www.gencourt.state.nh.us/rsa/html/XXXVII/420-G/420-G-11-a.htm (last accessed March 12, 2012). See also New Hampshire Admin. Rules, Parts Ins 4003-4004, http://gencourt.state.nh.us/rules/state_agencies/ins4000.html (last accessed March 12, 2012).

[16] APCD Council, *Standardization of Data Collection in All-Payer Claims Databases*, January 2011, pg. 1, http://apcdcouncil.org/sites/apcdcouncil.org/files/Standardization%20Fact%20Sheet_FINAL_for010711release_1.pdf.

[17] Keely Cofrin Allen, *Why All Payer Claims Databases are "all the rage" in health care*, Health IT Exchange, June 16, 2011, http://searchhealthit.techtarget.com/healthitexchange/CommunityBlog/why-all-payer-claims-databases-are-all-the-rage-in-health-care/.

Services (NHDHHS) designated Onpoint Health Data to collect and process data from plans.[18] Onpoint Health Data receives structured data from the plans, checks the submitted data for errors and completeness, harmonizes the data across carriers, and then populates New Hampshire's data warehouse with the validated data.[19]

Sample data flow diagram of centralized model:



Privacy and data security are frequently among the first considerations of organizers and policymakers when planning an APCD, and many privacy and security protections are explicitly required in state regulations that implement the APCD. Protections vary by state but fall under two general categories, often used in combination[20] –

- **Alterations of the data**: Due in part to HIPAA, which imposes fewer restrictions on the use of data that have been stripped of common identifiers, the most common protective alteration of APCD data used by states is stripping or camouflaging sensitive data elements via cryptography (such as encryption or hashing[21]).

[18] Onpoint Health Data, Clients, http://www.onpointhealthdata.org/about/clients.php (last accessed March 12, 2012).

[19] New Hampshire Comprehensive Health Care Information System, Frequently Asked Questions, http://www.nhchis.org/faq.html (last accessed February 16, 2012).

[20] National Association of Health Data Organizations, *Review of States' Confidential Data Release Policies and Recommendations for New York Statewide Planning and Research Cooperative System*, pg. 11, August 2008, (on file with author). See also, Barbara Rudolph, Gulzar Shah, and Denise Love, *Small Numbers, Disclosure Risk, Security, and Reliability Issues in Web-based Data Query Systems,* Journal of Public Health Management & Practice, Vol. 12 (2), March/April 2006, 176-183, http://journals.lww.com/jphmp/Abstract/2006/03000/Small_Numbers,_Disclosure_Risk,_Security,_and.10.aspx.

[21] A hash function takes a set of data and condenses it into a "representation" comprised of alphanumeric characters. Hashing two identical sets of data will produce identical representations. This technique can

- **Administrative controls**: Data use agreements (DUAs) are frequently utilized as administrative controls. The protections are typically correlated with the sensitivity of the data being released. For example, New Hampshire regulations require NHDHHS to enter into an agreement with its designated third party data collector/processor that "strictly" prohibits the designee from collecting or releasing data that contain patients' direct identifiers.[22] As a result, NHDHHS' designee, Onpoint, requires plans to hash direct identifiers (such as name, SSN, email address, personal photographs and biometric identifiers) prior to transmitting the data to Onpoint via secure web portal.[23]

APCDs use the data they collect for broad-based evaluations and improvements to health care delivery and population health. The types of analyses many APCDs run on their data are similar: most APCDs analyze for cost, efficiency, and quality of care across geographic regions, plans, and facilities.[24] Some state agencies assemble this data –stripped of individual identifiers – into products available to the public.[25] Some APCDs provide private and government researchers with data sets of varying detail, subject to data use agreements. The researchers use this data for a wide variety of projects, such as cost comparison analysis, disease prevalence, developing quality measures, and assessing barriers to care for particular patient populations. The data products furnished by New Hampshire's APCD are broken into three general categories, containing varying levels of detail and subject to different rules. State regulations require New Hampshire's APCD to make available on request "public use data sets," "limited use health care claims research sets," and "confidential health care claims research data sets."[26]

- **Public use data sets** contain no direct or indirect identifiers for individuals, health care practitioners, employers or purchaser groups. NHDHHS is required to release public use

---

therefore be used to match identical pieces of data without actually viewing the underlying data. See National Institute of Standards and Technology, Computer Security Resource Center, Cryptographic Hash Project, http://csrc.nist.gov/groups/ST/hash/index.html (last accessed February 15, 2012).

[22] New Hampshire Admin. Rules Part Ins 4004.01(g), http://gencourt.state.nh.us/rules/state_agencies/ins4000.html (last accessed February 16, 2012).

[23] New Hampshire Admin. Rules Parts Ins 4002.01(q), 4004.02, http://gencourt.state.nh.us/rules/state_agencies/ins4000.html (last accessed February 16, 2012). See also New Hampshire Comprehensive Health Care Information System, Frequently Asked Questions, http://www.nhchis.org/faq.html (last accessed February 16, 2012).

[24] Denise Love, William Custer, and Patrick Miller, All-Payer Claims Databases: State Initiatives to Improve Health Care Transparency, The Commonwealth Fund, September 2010, pgs.4-5, http://www.commonwealthfund.org/~/media/Files/Publications/Issue%20Brief/2010/Sep/1439_Love_allpayer_claims_databases_ib_v2.pdf. See also, http://www.apcdcouncil.org/state/new-hampshire

[25] See, e.g., New Hampshire HealthCost, www.nhhealthcost.org, (last accessed March 12, 2012). NH HealthCost disseminates information on the price of services offered by different providers.

[26] New Hampshire Admin. Rules Parts He-W 950.04-950.07, http://www.gencourt.state.nh.us/rules/state_agencies/he-w900.html (last accessed February 16, 2012).

data sets upon request. NHDHHS is required to maintain a record of releases of these data sets; the record is available for public inspection on NHDHHS' website.[27]

- **Limited use health care claims research data sets** hash all direct patient identifiers and insured group or policy numbers. If approved by a data release advisory committee, NHDHHS must release the limited use health care research data set on written request. The written request must contain a description of the research protocol to be performed and the procedures that will preserve the confidentiality of the data (such as indirect patient identifiers). Researchers using the data are also subject to a DUA with NHDHHS that requires safeguards to preserve confidentiality, prohibits re-identifying patients or employer/purchaser groups, requires the researcher to gain final approval before the research product is publicly released to ensure DUA compliance, and requires the researcher to return or destroy the data set upon completion of the project.[28] NHDHHS keeps a public log of limited use data requests on its website.[29]

- **Confidential research data sets** contain patient identifiers and are used with the informed consent of the identified patients. If the data set contains group policy identifiers, the set is used with the informed consent of the identified insurer groups. A privacy review committee must approve the use of a confidential research data set. As with the limited use health care claims research data set, the requestor of the confidential data set must make the request in writing with a description of the research protocol to be performed and the procedures that will preserve the confidentiality of the data. Researchers using the data are also subject to a DUA with NHDHHS similar to the DUA required of researchers seeking to use the limited use health care claims research data set.[30] Confidential research data sets are rarely, if ever, released to entities outside of NHDHHS.

III. Data centralization raises privacy and security risks

The goals and purposes of most secondary use programs are largely positive. Although many secondary use programs have taken steps to integrate privacy controls in the collection and use of sensitive data, a fundamental problem with the centralized architecture used by many secondary use programs is that centralization does not minimize the copying of data. Instead, it typically necessitates the maintenance and sharing of multiple copies of patient data – not only

[27] New Hampshire Admin. Rules Parts He-W 950.04-950.05, http://www.gencourt.state.nh.us/rules/state_agencies/he-w900.html (last accessed February 16, 2012). See also New Hampshire Comprehensive Health Care Information System, Public-Use Data Requests, http://www.nhchis.org/ (last accessed February 16, 2012).

[28] New Hampshire Admin. Rules Part He-W 950.06, http://www.gencourt.state.nh.us/rules/state_agencies/he-w900.html (last accessed February 16, 2012).

[29] New Hampshire Comprehensive Health Care Information System, Limited Use Data Requests, http://www.nhchis.org/ (last accessed February 16, 2012).

[30] New Hampshire Admin. Rules Part He-W 950.07, http://www.gencourt.state.nh.us/rules/state_agencies/he-w900.html (last accessed February 16, 2012).

is there the source copy and the copy with the agency operating the secondary use program, but also researchers or other third parties often receive their own copies of the data for their analytic functions. This pattern repeats itself each time a new research or policy need requires the creation of another centralized database. Yet continually building and copying huge repositories of medical data is risky, inefficient, and a poor long-term strategy:

- **Data breaches:** Maintaining copies of sensitive information in various locations for long periods of time sharply worsens the risk and severity of data breaches. Breaches of identifiable medical data are a growing – and extremely costly – problem for patients, health care companies, and government agencies.[31] Even if the data is de-identified, there is still some risk – albeit much lower – associated with breach and misuse.[32]

- **Public trust:** Unnecessarily funneling copies of patients' identifiable data to state and federal agencies for purposes other than treatment or payment often inflames public perception of government snooping, eroding both trust in the confidentiality of medical records and support for health care reform.[33] As HHS has stated many times, public trust in the privacy of digital health records is fundamental to the evolution to a modern, information-driven health care system.[34] Good health care depends on good information, but studies regularly show that patients who do not trust the confidentiality of their data are much less likely to be open with their care providers – sometimes withholding important information to preserve their privacy.[35]

---

[31] Ponemon Institute and ID Experts, *Benchmark Study on Patient Privacy and Data Security*, November 9, 2010, http://www2.idexpertscorp.com/press/healthcare-news/new-ponemon-institute-study-finds-data-breaches-cost-hospitals-6-billion.

[32] See Center for Democracy & Technology, *Encouraging the Use of, and Rethinking Protections for De-Identified (and "Anonymized") Health Data*, June 2009, pgs. 7-8, http://cdt.org/healthprivacy/20090625_deidentify.pdf.

[33] See Rep. Tim Huelskamp, *Obamacare HHS rule would give government everybody's health records*, September 23, 2011, http://washingtonexaminer.com/opinion/op-eds/2011/09/obamare-hhs-rule-would-give-government-everybody-s-health-records. See also Rep. Denny Rehberg, *Chairman Rehberg Investigates Possible Violations of Private Health Care Information Under President Obama's Health Care Plan*, October 13, 2011, http://pressrehberg.congressnewsletter.net/mail/util.cfm?gpiv=2100078808.1461.269&gen=1.

[34] David Blumenthal and Georgina Verdugo, *Building Trust in Health Information Exchange*, *Statement on Privacy and Security*, U.S. Dept. of Health and Human Services, http://healthit.hhs.gov/portal/server.pt?CommunityID=2994&spaceID=11&parentname=CommunityEditor&control=SetCommunity&parentid=9&in_hi_userid=11673&PageID=0&space=CommunityPage (last updated July 8, 2010).

[35] See Markle Foundation, *Common Framework for Private and Secure Information Exchange, The Architecture for Privacy in a Networked Health Information Environment*, April 2006, pgs. 3-4, http://www.markle.org/sites/default/files/P1_CFH_Architecture.pdf. In a recent study, more than a quarter of U.S. patients stated they would withhold information from clinicians and avoid treatment in order to preserve the confidentiality of their health data. New London Consulting and FairWarning, *UK: How Privacy Considerations Drive Patient Decisions and Impact Patient Care Outcomes*, pg. 11, October 6, 2011, http://www.fairwarningaudit.com/documents/2011-whitepaper-uk-patient-survey.pdf.

- **Inefficient, costly and burdensome**: Diverse entities at the state and federal level want access to health data for secondary use programs that sometimes have very similar goals.[36] It is burdensome and costly for plans to set up and secure multiple large data submissions to different entities in various locations, especially if those entities require different data formats. In addition, it is costly for society as a whole when numerous government entities establish and maintain multiple large centralized databases. This situation is particularly inefficient when the entities are performing substantially similar analyses.

- **Scope creep**: When government possesses copies of health data, there is a risk that the government will incrementally expand its uses of the data beyond the limited set of purposes described when the program was established. While the public may have participated, directly or through their elected representatives, in the processes that originally created the databases, they may have few or no meaningful opportunities to learn about, comment upon, and vet new uses of data already in government possession.

IV. Decentralized alternatives

Instead of requiring the creation of yet more centralized databases stocked by data feeds from health plans, policymakers should consider decentralized alternatives. Distributed networks support the coordination of multiple, autonomous databases to meet a shared objective – such as analyzing database content for research purposes – without requiring the creation of a central data repository.[37] Distributed networks can often cost less and take less time to establish than centralized databases because a distributed network minimizes data transfer and leverages existing infrastructure – such as databases, security safeguards, and human capital. If many more initiatives analyzing health data are launched, which is highly likely, it would be less burdensome for plans to set aside one copy of their data and manage secure access than to set up separate data feeds with multiple agencies. Furthermore, keeping most data with the original data sources may help ease some of the proprietary and liability concerns that many data sources have with regularly transferring whole copies of large data sets to the government. Using a distributed network can also reduce the risk and severity of data breaches compared to centralized databases. Leaving data sets with the original data sources minimizes the number of copies of sensitive data sets in circulation. Fewer copies of sensitive data about individuals (i.e., compiled from multiple databases) means that not only are less data overall leaked in the event of a breach, but less data about each affected individual are subject to inappropriate exposure.

---

[36] For example, OPM's Health Claims Data Warehouse and many state APCDs perform cost and quality comparisons across geography and demographics.

[37] Brown et al., *Design Specifications for Network Prototype and Cooperative To Conduct Population-Based Studies and Safety Surveillance*, Effective Health Care Research Report Number 13, Agency for Healthcare Research and Quality, July 2009, pg. 3, http://www.effectivehealthcare.ahrq.gov/ehc/products/54/150/2009_0728DEcIDE_DesignSpecNetCoopPopSafety.pdf.

There are multiple approaches to decentralized analytical systems, but not all are equally appropriate for every secondary use program – the best fit will depend on resource constraints and the type of analytics required. Policymakers should consider which model could achieve their research goals while maximizing data security and accuracy. In some cases, as with the proposed Multi-Payer Claims Database (MPCD), the architecture may be a hybrid combining more than one model.[38]

One decentralized approach is for researchers to send payers detailed research questions, permitting payers to write analytic code to answer those questions. The payers use the code they have written to analyze their in-house data and then return structured responses – rather than copies of the data – to the researchers. Payers are not required to maintain data in a common format for this process to work. An operational example of this type of distributed system is the Federal Partners project of the Food and Drug Administration (FDA).[39] This *"distributed query"* approach might be suitable for many common secondary uses, such as the research goals envisioned by many APCDs – cost comparison analysis, developing quality measures, measuring disease prevalence, and identifying barriers to care access. Nonetheless, this query-based approach may be inappropriate for secondary uses that can lead to competitive advantages or disadvantages for health plans, such as CMS' risk adjustment program.[40] Permitting plans to write the code and analyze their own data for risk adjustment could make the risk adjustment program susceptible to inaccuracy and fraud; smaller insurers could make errors due to a lack of experience in conducting risk adjustment analysis, while others might intentionally submit falsified results to gain a favorable outcome – and CMS might have to rely on retroactive audits of outdated, untrustworthy data to gauge the accuracy of the data used for the secondary use program.[41]

A second query-based approach to decentralized analytics is for researchers to write the analytic code and send the code to payers. Payers analyze their in-house data using the code (but do not modify the code), review the output, and provide the responses to the research questions with computer logs that reveal any manipulation of the code. This process does require the payers to use a common data format. This decentralized approach greatly reduces

---

[38] The MPCD will access longitudinal claims data (and eventually information from EMRs) for comparative effectiveness research. Though it is early in the implementation process, the proposed MPCD architecture would centralize health information of lower sensitivity into a database, but leave sensitive and identifiable information with the health plans, where it would be made accessible via a distributed system. See U.S. Dept. of Health and Human Services, Multi-Payer Claims Database (MPCD) for Comparative Effectiveness Research, June 16, 2011, pg. 6, http://www.ncvhs.hhs.gov/110616p1.pdf.

[39] U.S. Food and Drug Administration, *The Sentinel Initiative*, July 2010, pg. 6, http://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM233360.pdf.

[40] 76 Fed. Reg. 41930, 41940-1.

[41] Edwin Park, *Allowing Insurers to Withhold Data on Enrollees' Health Status Could Undermine Key Part of Health Reform*, Center on Budget and Policy Priorities, Dec. 12, 2011, pgs. 3-4, http://www.cbpp.org/files/12-12-11health.pdf. See also Center for Democracy & Technology, Comments to CMS–9975–P, October 31, 2012, pgs. 5-6, http://cdt.org/files/pdfs/CDT_Comments_to_CMS-9975-P.pdf

the risk of fraud by prohibiting the payers themselves from writing or manipulating the analytic code, although payers themselves still analyze their own data when responding to queries. An operational example of this approach is the FDA's Mini-Sentinel Initiative. Sentinel was launched in 2008 in order to quickly monitor the safety of products the FDA regulates. Mini-Sentinel provides a secure web interface through which users authorized by the FDA can query product data and send questions to the data sources (which include health plans), but the data remain with and are managed by the participating data sources.[42]

In CMS' 2011 proposed rule for its risk adjustment program, the agency expressed its concerns regarding distributed query systems. Although CMS acknowledged the potential privacy benefits of distributed systems, the agency was concerned that permitting plans to analyze their own data could lead to fraud and inaccuracy, and CMS was also concerned about small insurers' ability to respond to multiple queries.[43] These issues drove CMS to propose regulations that would have locked plans participating in the program into a centralized database model whereby the plans would have submitted claims data to a centralized system operated either by CMS or individual states.[44] However, in its final rule, issued in March 2012, CMS changed course.[45] In the final rule, CMS established that it would use a distributed system when it analyzes data for risk adjustment, while states would have the flexibility to choose the data collection model that worked best for them, centralized or not.[46] While CMS' final rule notes that plans participating in the risk adjustment program would need to maintain their data according to HHS requirements, the final rule gives little detail regarding what a distributed system for risk adjustment might look like or how the distributed system CMS envisions would overcome the agency's concerns regarding fraud, inaccuracy.

For secondary use programs that require person-level data and carry an unacceptable level of risk of fraud or inaccuracy, CDT recommends policymakers explore a "*distributed access*" model. The distributed access model would give agencies direct access to (de-identified) data and permit the agencies – not the plans – to perform the analyses.[47] In the following paragraphs, we describe the distributed access model in more detail.

---

[42] Sentinel also operates under established privacy and security standards aimed at constant protection of personal information. See U.S. Food and Drug Administration, *The Sentinel Initiative: National Strategy for Monitoring Medical Product Safety*, May 2008, pgs. 13, 15, http://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM124701.pdf.
Additional examples of operational query systems include the Observational Medical Outcomes Partnership, the Vaccine Safety Datalink, and the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) program.

[43] 76 Fed. Reg. 41940-1.

[44] *Id.*

[45] 77 Fed. Reg. 17220, 17233. See also Center for Democracy & Technology, Comments to CMS–9975–P, October 31, 2012, http://cdt.org/files/pdfs/CDT_Comments_to_CMS-9975-P.pdf.

[46] *Id.*

[47] Similar systems utilizing federated architecture are already in use in other sectors. See Comments of Palantir Technologies to CMS proposed rulemaking on Standards Related to Reinsurance, Risk Corridors and Risk Adjustment, CMS–9975–P, 76 Fed. Reg. 41930, October 31, 2012, pgs. 2-3, available at

Under this model, each health plan participating in a secondary use program would be required to set aside a structured, de-identified copy of its claims and encounter data in a secure environment (such as on an edge server or in a cloud storage center).[48] The data in the secure environment should be in a uniform format according to government standards.[49] Each plan would offer an interface over the Internet to the state and federal agencies responsible for operating the secondary use programs, providing secure access to the data set aside in the payers' respective systems.[50] The interface for the distributed access model we propose would need to be flexible enough to handle a wide variety of queries, although access controls can restrict the queries a particular user is permitted to make. The agencies themselves would use this access to perform the analyses necessary to meet the goals of their secondary use programs, but the data would not be duplicated and sent to the government, and agencies would be prohibited from using the data for anything other than the specified uses of the programs.[51] The agencies would retain the results of their analyses, but would not keep full copies of the data.[52]

Under a distributed access model, participating plans would be required to de-identify the data held "at rest" in the secure environments.[53] To the extent that state or federal agencies need longitudinal records of individual patients, plans could use a one-way hash algorithm to mask patient identifiers while allowing agencies to track records belonging to the same patient.[54] To

---

http://regulations.gov (last accessed February 15, 2012).

[48] Ideally, the secondary use program would have clear data requirements and plans would only have to populate the secure environment with the minimum quantities of data necessary to satisfy the program needs.

[49] Many state APCDs already require (through regulation) providers and plans to standardize data elements prior to submission to their designated data processors.

[50] The FDA's Mini-Sentinel uses software that provides a secure web interface to submit and receive query programs and analytical results. See Platt et al., *The U.S. Food and Drug Administration's Mini-Sentinel program: Status and Direction*, Pharmacoepidemiology and Drug Safety 2012; 21(S1): 4, available at Wiley Online Library DOI: 10.1002/pds.234, http://onlinelibrary.wiley.com/doi/10.1002/pds.2343/abstract.

[51] Plans that exit a secondary use program may be required to maintain the secure system for a time period sufficient to enable the agencies to complete any data analysis necessary to meet program needs. Center for Democracy & Technology, Comments to CMS–9975–P, October 31, 2012, pg. 5 http://cdt.org/files/pdfs/CDT_Comments_to_CMS-9975-P.pdf.

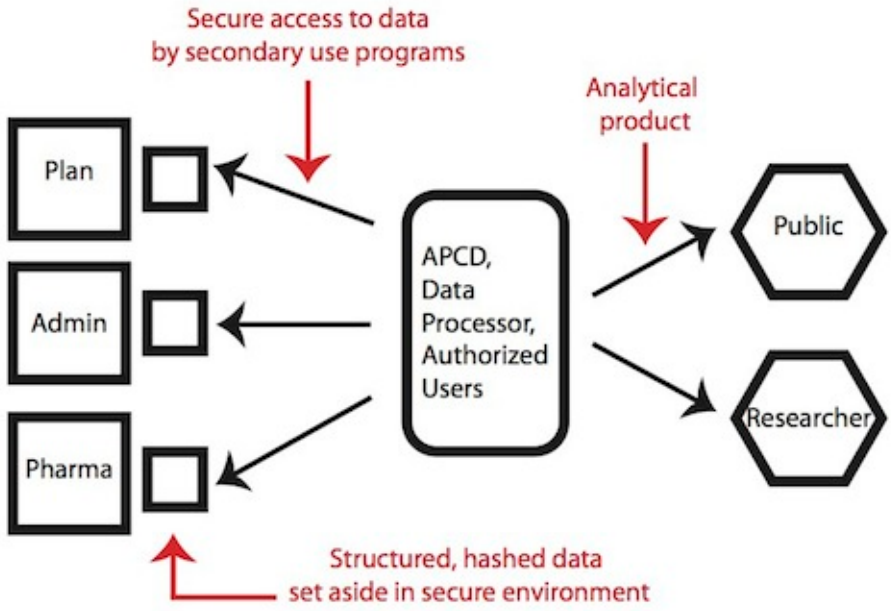
[52] Researchers using a distributed access system could still release their consolidated results to the public or other parties, as some APCDs do now using centralized databases.

[53] 45 CFR 145.514(b).

[54] In order to match individual patients held in different recordkeeping systems – such as across plans or care settings – each system likely must hash the patients' identifiers with the same algorithm.

maintain the protective effect of the hash, it would be crucial for plans and agencies to exercise appropriate key management.[55]

Sample data flow diagram of distributed access model:



Although the secure environment would operate adjacent to the databases plans use for normal business operations, strong firewalls could separate the two systems and plans could upload data to the secure environment periodically (rather than "in real time"). This would help prevent unauthorized parties from using the secure environment to gain access to the plans' proprietary databases. Plans could de-identify the data prior to uploading them to the secure environment. Plans would need to work closely with agencies to establish clear lines of responsibility for data breaches and unauthorized access. Researchers and agency employees seeking access to plans' secure environments would need to be credentialed and authenticated.

Ideally, the distributed access system would facilitate different tiers of access, rather than offering a choice between no access or full access. Identity and access management software could grant access to the secure environment based on the program or the role of the user. While plans may be required to give agency-run secondary use programs virtually unrestricted access to the secure environment, plans should not be required to do so for all research activities. The access management software could give plans or agencies the ability to review and approve or deny requests for access to the secure environment.[56] The tiered access

[55] See National Institute of Standards and Technology, Computer Security Resource Center, Key Management, http://csrc.nist.gov/groups/ST/toolkit/key_management.html (last accessed February 15, 2012).

[56] The Massachusetts eHealth Institute (the state-designated health information exchange of Massachusetts) currently uses software, called PopMedNet, with this capability. PopMedNet enables external researchers to analyze and retrieve data held in multiple locations, subject to the approval of the data sources. Massachusetts eHealth institute, MDPHnet - Distributed Data Analytics,

scheme could be flexible enough to enable secondary use programs to grant researchers selective, secure access to health data – just as New Hampshire's Comprehensive Health Care Information System releases "public use," "limited use," and "confidential" data sets to support vetted research requests.

By enabling agencies to perform their own calculations, and because the data held in the secure environment would be formatted according to government standards, the distributed access system would reduce the risk of plans providing inaccurate data to the risk adjustment program. As with any system collecting and analyzing individual-level health data, the distributed access system proposed above should also incorporate policies and technical mechanisms that hold health plans accountable for the reliability of health data held in the secure environment, such as immutable audit trails recording actions performed on the health data.[57] Agencies could require plans to certify the accuracy of the data they submit and subject plans to penalties for chronic or willful failure to meet accuracy standards. Agencies could periodically audit the data submitted to the plans' secure environments for accuracy, comparing the plans' submissions to normative data and matching the submissions to the plans' internal records.

The distributed access approach faces several important technical challenges, primarily related to frequency of access and network reliability. For example, a distributed access system will need to ensure the stability of data content over multiple sessions in high-availability servers, especially as the quantity of access requests grows with the number of new secondary use programs. A distributed access system will also need to maximize the reliability of a decentralized network of non-redundant access points. Some of these issues might be mitigated by assigning plans specific windows of time in which their secure environments must be accessible, rather than requiring the data to be available at all times. These and other challenges must be evaluated and addressed prior to full implementation of a distributed access system.

V. Conclusion

Ultimately, any decentralized solution must effectively support secondary use programs, and no system should be formally deployed until its functionality is validated. However, now is the time to consider the network architectures for secondary use programs and health information exchange, when the technical infrastructure is in a relatively nascent stage and before too many regulatory mandates bar any non-centralized options. CDT believes it is imprudent for federal and state agencies to issue regulations that lock agencies and plans into the centralized model of health data analysis for secondary use programs. Instead, regulations should leave open the possibility for plans or states to use decentralized models, subject to the approval of federal and state agencies. That way, the agencies need not approve any architecture that does not adequately support their program goals, but at least the option to adopt a decentralized solution will be available in the future without requiring significant regulatory modifications.

---

http://www.maehi.org/what-we-do/hie/mdphnet (last accessed Feb. 6, 2012).

[57] Health plans might also use audit trails to validate and verify – but not modify – the results of researchers' and agencies' analyses.

CDT urges HHS to develop a strategy, in collaboration with technology vendors, state agencies, and consumer groups, to comprehensively explore models of analysis that do not require health data to be copied and stored in multiple databases. CDT is encouraged by HHS' existing efforts to explore decentralized systems that are focused on discrete program objectives – such as the Mini-Sentinel and Federal Partners projects – as well as CMS' decision to support a decentralized approach for its risk adjustment program.[58] CDT urges HHS to initiate projects to assess the effectiveness of distributed systems for broader, population-scale health data analytics using de-identified data.

The immediate goal of our recommendations is to ensure that regulations governing secondary use programs give government agencies and health plans greater flexibility to use distributed systems in the future, rather than requiring purely centralized solutions. The long-term goal is to leverage pilot projects to establish a scalable, secure architecture that effectively supports valuable secondary use programs while minimizing unnecessary duplication and transmission of patients' sensitive data.

For further information, please contact Harley Geiger: harley@cdt.org, 202-637-9800.

[58] 77 Fed. Reg. 17220, 17233.