MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-584

Patient-Centered Outcomes Research in Practice: The CAPriCORN Infrastructure

Anthony Solomonides[a], Satyender Goel[b], Denise Hynes[c,l], Jonathan C Silverstein[a], Bala Hota[d1], William Trick[e], Francisco Angulo[e], Ron Price[f], Eugene Sadhu[c], Susan Zelisko[f], James Fischer[c], Brian Furner[g2], Andrew Hamilton[h], Jasmin Phua[i], Wendy Brown[j], Samuel F Hohmann[k,d2], David Meltzer[g1], Elizabeth Tarlov[c,l], Frances M Weaver[f,l], Helen Zhang[e], Thomas Concannon[m], Abel Kho[b]

[a] Center for Biomedical Research Informatics, NorthShore University HealthSystem
[b] Feinberg School of Medicine, Northwestern University
[c] University of Illinois, Chicago
[d]Rush University Medical Center—[1]Department of Medicine, —[2]Dept of Health Systems Management
[e] Cook County Health and Hospital Systems
[f] Loyola University Health System
[g] University of Chicago—[1]Medicine, —[2]Center for Research Informatics
[h] Alliance of Chicago Community Health Services
[i] Medical Research Analytics and Informatics Alliance
[j] VA Jesse Brown Hospital
[k] Universities Healthsystem Consortium
[l] VA Edward Hines Hospital
[m] RAND Corporation.

Abstract

CAPriCORN, the Chicago Area Patient Centered Outcomes Research Network, is one of the eleven PCORI-funded Clinical Data Research Networks. A collaboration of six academic medical centers, a Chicago public hospital, two VA hospitals and a network of federally qualified health centers, CAPriCORN addresses the needs of a diverse community and overlapping populations. To capture complete medical records without compromising patient privacy and confidentiality, the network created policies and mechanisms for patient consultation, central IRB approval, de-identification, de-duplication, and integration of patient data by study cohort, randomization and sampling, re-identification for consent by providers and patients, and communication with patients to elicit patient-reported outcomes through validated instruments. The paper describes these policies and mechanisms and discusses two case studies to prove the feasibility and effectiveness of the network.

Keywords:

Patient-Centered Outcomes Research; Comparative Effectiveness Research; Electronic Health Records; Data Collection; Data Linkage; Aggregation; Data Sets; Deidentification;Re-identification; Consent.

Introduction

PCOR, CER and PCORnet

The Patient-Centered Outcomes Research Institute (PCORI) was established following the US Patient Protection and Affordable Care Act in 2010. PCORI's mission is to advance and support Patient-Centered Outcomes Research (PCOR), which "helps people and their caregivers communicate and make informed healthcare decisions, allowing their voices to be heard in assessing the value of healthcare options." [1]

In particular, PCOR:

- Encompasses comparative effectiveness research (CER) on interventions to inform decision making.
- Addresses individuals' (especially patients' and care-givers') preferences and autonomy.
- Studies a diversity of settings and populations.
- Seeks to balance stakeholders' concerns, including burden to individuals and availability of resources.

One principal action of PCORI is to support 11 Clinical Data Research Networks (CDRN) and 18 Patient-Powered Research Networks (PPRN). Both kinds of research networks are seen as infrastructure-building projects, with specific structural, process and outcome goals to prove the feasibility and usefulness of the networks. CDRNs focus on major academic medical centers: apart from demonstration of viable infrastructures, CDRNs demonstrate their value by conducting research in a number of specific conditions. Each network nominates the conditions on which it will work. However, longer term sustainability for the infrastructure can only be achieved through success in early studies, and proving to the research community that the network represents a valuable resource that is worth both exploiting and supporting through further funded studies and grant proposals. PPRNs focus on specific conditions that are of concern to patients, care providers, and patient advocacy organizations. Many networks have formed around existing formal or informal networks of support and advocacy groups.

Overarching the CDRNs and PPRNs, PCORI established a supra-network, PCORnet, that acts as a collaboration venue, clearing house, and policy-development body. Best conceived of as a network of networks, PCORnet ensures that the infrastructures created by the different CDRNs and PPRNs will remain interoperable and responsive both to researchers' needs and to the expectations of patients, care providers and advocates.

CAPriCORN

One of the CDRNs, CAPriCORN, represents an alliance of Chicago institutions collaborating in recognition of the need for pre-competitive comparative effectiveness research (CER) in their highly diverse community—diverse both in the type of

institutions involved and, importantly, in the populations they serve. CAPriCORN is not typical of CDRNs, although it shares many characteristics. Some of its unique features provide a model for collaboration in environments where, for example, patient populations at different institutions overlap, where nevertheless a full picture of each patient's health record is necessary for meaningful research results.

Data for sharing within CAPriCORN—and in the wider community at a later stage—will be in a HIPAA-compliant, de-identified format. Two working groups (WG), Informatics WG and Ethics and Regulatory WG, devised a federated data architecture, a data model with appropriate standards, and a designed data flow engineered to ensure that no protected health information (PHI) is released other than under strictly controlled conditions and, at the same time, maintaining the research value of the data that is released. De-identified data will be released on a study-by-study basis. A statistically benchmarked process is used to generate a pseudonymous identity for each patient in such a way that distributed patients' records across different providers in the network can be matched and integrated. The records are not brought together into a single central database, but are instead put in a virtual repository – by allowing distributed queries across the different systems through the validated mechanism of PopMedNet [4, 5]. Consent will be sought when access to PHI, or directly to the patient for patient-reported outcomes, is necessary.

Methods

Population

CAPriCORN comprises a network of six academic medical centers (University of Chicago, University of Illinois, Chicago, Loyola University, NorthShore University HealthSystem, Northwestern University and Rush University Health), the Alliance of Chicago's Federally Qualified Health Centers, a major public hospital, Cook County Hospital, and two Veterans Affairs hospitals, VA Edward Hines and VA Jesse Brown. Geographically, these institutions serve the greater Chicago metropolitan area and are available to a total population of approximately 9.5 million. (In addition to these "data-providing" institutions, 22 other organizations contribute research, patient advocacy, and infrastructure services to CAPriCORN. Their role is described below.)

CAPriCORN institutions together held 2,860,000 covered lives in electronic health records. A preliminary analysis of seven of the ten institutions indicated 6,923,111 patients, of whom 1,465,285 were registered with a primary care provider; however, after de-duplication, the numbers were 5,741,268 and 1,242,380 unique patients respectively. Thus some 20.6% of patients are associated with more than one institution, and even among the primary populations, there are 18% of patients with more than one PCP registration. This appears to be symptomatic of deprivation in the inner city, where economic necessity requires individuals to move opportunistically from provider to provider.

The racial breakdown of the primary population is 47.5% Caucasian, 27.9% African American and 14.9 Hispanic, with just over 9% in other categories. Of this population, 59.3% are female, 40.7% male. The mean age is 50 with a standard deviation of 17.9.

De-identification and De-duplication

While fragmented care may be suboptimal, research on comparative effectiveness of treatments requires as accurate and as complete a record of each patient's health status and episodes of illness as can be reconstructed, if meaningful and valid results are to be achieved. With multiple records for up to 20% of patients, de-duplication is strongly indicated. The means of achieving this lie in a method of de-identification.

In the US, there is currently little prospect of a single unique patient identification code. Where health information exchanges have been instituted, it is necessary to implement an "enterprise master patient index" (EMPI), but even these are rare because of a number of concerns, principally privacy and security, and economics and sustainability. Nevertheless, prior experience was sufficiently encouraging to suggest that a specific design and implementation in the Chicago area would be worthwhile. This prior knowledge and experience provided a fundamental cornerstone for the CAPriCORN network.

The de-identification algorithm comes from Kho *et al* [2, 3]. The algorithm uses a set of strictly personal identifiers, i.e., nothing that may be institution-specific, to generate up to 17 different combination strings and uses a statistically selected subset of these to construct a "hash-ID." The hashing algorithm is not reversible, but its high specificity allows patients who have multiple records to be discovered, albeit anonymously.

Organizational Design

CAPriCORN is led by a Principal Investigator at the Chicago Community Trust, an organization focused on civic leadership and philanthropy. A Steering Committee is the decision-making body, whose composition was designed around the natural concerns of a network to conduct and facilitate patient-centered outcomes and comparative effectiveness research across a number of healthcare institutions. The Steering Committee also reflects the underlying architectural design of the infrastructure and the projected governance and regulatory framework of that infrastructure.

Clinical Data Research Networks are intended to be open to external collaboration, explicitly designed to be open to patient concerns, and subject to all the normal ethical and regulatory processes that apply to human subjects and social science research. These are, respectively, reflected in the network's External Researcher Committee, Patient and Clinician Advisory Committee, and Chicago Area Institutional Review Board (CHAIRb). All these committees define processes and workflows for patient and carer consultation, the triage of internal and external research proposals, the handling of data requests, the release of data, and the consenting process prior to any re-identification of and contact with patients.

Critical to the infrastructural design are two "honest broker" roles in the network. Other than in very specific, precisely defined circumstances involving only consented patients, these organizations hold no protected personal health information (PHI) but handle the "de-identifiers", principally the hash-IDs for de-duplication, and subsequent to the definition of specific condition cohorts, a second level of pseudonymization, the cluster-IDs, which are randomly generated "per study, per hash-ID" thus avoiding any unintended crosstalk between independent studies.

The principles, explicit and implicit, that guided this design are:

- All studies, including those submitted as "proof of principle" for the network, along with new and external proposals, will be subject to triage by the Patient and Clinician Advisory and External Researcher committees, then subject to review by CHAIRb, with the ultimate decision resting with the Steering Committee.

- All PHI will be held at institutions, benefiting from all the protections (firewalls, authorizations, etc.) that each applies to its own patient data.
- The data collected will be strictly non-PHI and minimal with respect to any cohort identification needs (all that is needed, but no more).
- Identifiers will be hashed into pseudonymous "hash-IDs" for the purpose of de-duplication. Honest Broker 1 (HB1) will provide institutions with a unique "hash seed" that each will use to de-identify its own patients through hashing.
- The second honest broker, HB2, will use the hash-IDs provided by institutions to identify "duplication" and determine the set of institutions to which each patient corresponds. HB2 then generates a random identifier, the cluster-ID, for each unique patient in the given cohort. At this point, if considered necessary, the institutions themselves may be pseudonymized. (No PHI will flow to HB2.)
- Patients' records may only be linked through the hash-ID. Cohort identification for specific studies and non-PHI data requests from sites for constructing aggregate records may be conducted only by means of a distributed query mechanism (currently, PopMedNet [4, 5]) which allows inspection and vetting of queries prior to execution and results from queries to be examined prior to release.
- All studies that require access to PHI must identify a co-investigator at each site.
- Provider consent to approach patients to consent for particular studies will be requested, and subsequent patient consent will be sought, according to institutional rules and norms.
- Randomization of patients for consent will be done anonymously both in respect to patients and institutions.

As noted above, these principles are visible in the organizational structure of the network, but they are also evident in the architectural design of the infrastructure.

Network Architecture

The architecture and processes represented by the various flows in this diagram are detailed in Figure 1.

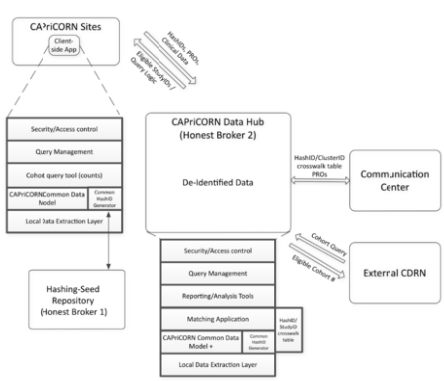

Figure 1 – A schematic diagram of the network displaying the two "honest broker" roles, the institutional repositories and the central "data hub" which hosts the matching and distributed query services.

CAPriCORN developed a data model and data standards, together with "extract-transform-load" processes for its institutional data marts. The data model is effectively based on a star schema with the concept *Encounter* at its center, so that data can be understood at a transactional level. A data dictionary was adopted showing domains and variables within them (apart from patient demographics, radiating out from encounters are diagnoses, medications, procedures, vital signs, laboratory results, and some additional local variables). Standards and terminologies indicate values in each category. The degree of privacy restriction for each variable (within-institution, within-CAPriCORN, within-PCORnet) is also indicated.

Each institution established a data mart (or other local database) which, notwithstanding the differences in platforms, precisely matches the CAPriCORN data model. Thus, although local adaptations of SQL queries will be necessary, the essential logic of queries submitted to the "data hub", i.e., the distributed query service, will remain unaltered, as required by PCORnet for its greater vision of seamless patient-centered, comparative effectiveness research.

A Communication Center is also being established to facilitate the process of re-identification of patients for provider consent to approach patients and for patient consent to participate in survey research (patient-reported outcomes, or PROs) and intervention studies. Each institution's processes are respected, and no pre-consent PHI flows through the center.

Process Description

1. HB1 hosts a stand-alone, generic hashing-seed generator application; it generates a SEED and passes it automatically to all participating institutions.
2. Each INSTITUTION uses the SEED and a set identifiers to generate a set of multiple *hashes* for each patient on record:

   [SSN, FirstName, LastName, DoB, Gender] ⊗ SEED
   → { hashes }

   from which a unique hash-ID is generated and cross-linked to the patient's MRN for internal identification.

This is per patient; […] signifies a vector of personal data.

Hash-IDs can be used within each INSTITUTION locally, if desired.

3. For each STUDY, every INSTITUTION runs the appropriate phenotyping algorithm to select its subpopulation of all unique patients who satisfy the cohort criteria. The hash-IDs along with all the hashes are returned to HB2.
4. For each study, HB2 collects all hashed data and de-duplicates, storing the result in a vector as follows:

   $\{ (\text{institutionID}=1) : \text{hash-ID}_1 \}$
   $\lozenge_{\text{hash-ID}} \ldots \lozenge_{\text{hash-ID}}$
   $\{ (\text{institutionID}=10) : \text{hash-ID}_{10} \}$
   → hash-ID : institutionVector

where $\lozenge_{\text{hash-ID}}$ represent the join on hash-ID. The patient's hash-ID and institutionVector now appear thus:

	Institutions									
Disease D	AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ

hash-ID										
xyz123	0	0	0	1	0	0	0	1	0	0

The patient whose hash-ID is "xyz123" was identified as having disease D and having partial records at UI and RU. We note that

(i) the hash-ID is in reality a more complex object (cf. [2]);

(ii) this may not be the complete record for this patient.

5. The five collections { hash-ID }, one for each study, are returned to all the institutions for cohort verification.

This is necessary, because, for example, a patient with an anemia record at one hospital (RU) may turn out to have a record at another hospital (UI) that does not mention anemia. Nevertheless, a complete record for that patient must include the partial records from both institutions.

6. Each institution checks the lists against its reference hash-ID list and so completes each patient's record if necessary.

For the sake of illustration, suppose now that we have found the patient above has also been seen at yet another hospital (CC) for an unrelated condition. The vector now becomes:

Disease D	AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ

hash-ID										
xyz123	0	1	0	1	0	0	0	1	0	0

We can now confidently compile a complete record of the patient.

7. At this point, HB2, as an honest broker, must do two more de-identification steps:

   a. disguise the institutions

   b. replace hash-IDs with non-derived ids for the patients; these are the cluster-IDs.

For the first step, HB2 randomly assigns pseudonyms to the institutions, say:

AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ
ff	*dd*	*aa*	*jj*	*bb*	*ii*	*cc*	*ee*	*hh*	*gg*

and these are then indexed as:

aa	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*
UC	LU	NU	CC	RU	AL	VJ	VH	NS	UI

The example patient now appears as:

Disease D	*aa*	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*

hash-ID										
xyz123	0	0	0	1	1	0	0	0	0	1

   c. The hash-IDs for each study cohort can now be replaced with unique cluster-IDs.

Our example patient now appears as:

Disease D	*aa*	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*

cluster-ID										
D-900093	0	0	0	1	1	0	0	0	0	1

Now, only possession of the table converting hash-IDs to cluster-IDs can enable anyone to re-identify the patient.

Distributed Queries

With cohort cluster-IDs collected, HB2 routes data requests through the distributed query service to the institutional data marts (IDMs). Locally, each institution will determine if the proposed query against its IDM is acceptable, allow the query to execute, and even then scrutinize the results before releasing them. Both in sending the requests and as results are received, HB2 can match cluster-IDs to hash-IDs, so that even a clinician researcher working on a project in their own specialty may be able to view expanded records of their own patients without recognizing them as their own. This provides a very high standard of de-identification.

Re-identification

Once particular studies based on entire cohorts are launched, re-identification of subsets of patients will most likely be necessary. Having received approval both from the Steering Committee (with advice from PCAC and ERC) and permission to proceed from CHAIRb, a researcher may request the Communication Center to randomly select a possibly weighted sample from across institutional or other populations for re-identification. The researcher will also be able to submit, through HB2, a data request for controls. Subject to CHAIRb's approval, institutional processes can be employed to gain provider consent and from there patient consent to participate in a study. Given the cluster-IDs of the patients in the study group, the Communication Center can alert institutions to the hash-IDs of patients to be approached for re-identification. In some cases, the Communication Center will also provide institutions with the means to collect patient-reported outcomes.

In the case of patients attending multiple institutions, which institution (or more precisely, which provider) should consent the patient for an identified study may be complex. A variety of algorithmic approaches is possible, including some that may work well but are computationally expensive. This may take the form of querying the system for the number of encounters at each institution in the last year (complex, but likely to reflect the patient's expectation) or it may suffice to look where the patient is registered for primary care (inexpensive, but may be irrelevant). The present ruling of CHAIRb only constrains the approach to be through a provider who is actually involved in the patient's care.

Results

Approximately at the halfway point in the project, achievements across a number of fronts include:

- Establishment of a sound governance structure, including a common central IRB, with data use and business associate agreements in place.
- Establishment and launch of a Patient and Clinician Advisory Committee with a clear role in the review, triage, and approval of new research proposals and a comprehensive manual for its operations.
- Approved design for the technological infrastructure, including a data model designed for ease of distributed query as well as with model evolution in mind.
- Approved processes and workflows now increasingly described and approved in protocols.

- Preliminary tests of the de-identification process and the distributed query machinery.
- Preliminary phenotyping in all five study cohorts proposed at project submission (see below). Preparatory phenotyping for a number of other studies, including incidental findings in osteoporosis, the national aspirin trial, bariatric surgery, antibiotics and childhood weight, bisphosphonates, and others.
- The de-identification and de-duplication processes in CAPriCORN are increasingly being reviewed as a model to be replicated across other CDRNs.

The internal organization of the network lends itself well to establishing CAPriCORN as a corporate entity; this would no doubt present new challenges, but is under consideration.

Discussion

The data model deployed at institutions to construct a data mart. Based on model variables, five phenotyping algorithms were devised and tested at multiple sites to identify overweight and obese patients (as required of all CDRNs); ambulatory patients suffering from asthma and in-patients with anemia (the two common disease cohorts); and patients with recurrent *Clostridium difficile* infection (RCDI) and sickle-cell disease sufferers (the two rare conditions).

In preparation for all these studies (and other anticipated future studies, including the PCORnet-inspired Aspirin trial and various collaborations with other CDRNs and PPRNs) the central IRB, CHAIRb, reviewed a Master Protocol which serves as a prefix to all specific study protocols.

Extract-Transform-Load (ETL) processes were undertaken against a number of different proprietory EHR systems. Some of these were shared publicly (e.g., through an EHR vendor's community sharing portal, thus conforming with requirements of commercial confidentiality). ETL logic was shared among all data-contributing sites to ensure compatibility.

The CAPriCORN data model is a superset of the PCORnet common data model against which external requests will be formulated. This model produces a straightforward mapping of data and requests from PCORnet to CAPriCORN. Additional data models influence the central PCORnet design, such as (Mini-)Sentinel, OMOP, i2b2 and others, and studied with a view to establishing correspondences should collaboration make a translation between CAPriCORN and another data model desirable.

Among the proposed cohort studies, the case of RCDI provides a convenient example of a hard test-case for the infrastructure. The study has not yet been completed, but based on data stored according to the data model and addressing queries to pre-existing institutional data warehouses rather than the institutional data marts, accurate cohort counts have been achieved.

Index cases of *Clostridium difficile* (CDiff) infection were identified, either by the presence of a diagnosis code or by laboratory test results. The first difficulty arises in recognizing resolved CDiff infection: how to differentiate between refractory and recurrent infection. If there is no encounter with CDiff code, laboratory test or relevant medication within eighteen days of date of diagnosis or of positive test result, the infection is assumed to have cleared. Any further infection in 18 to 56 days post index date is recorded as recurrence. Infections later than 56 days are considered new rather than recurrent.

One of the key challenges to CAPriCORN's distributed architecture will be the identification of recurrence across institutions. This challenge has not yet been attempted, but will be among the first studies that the system will address. The cohort is anticipated to be relatively small and the patient cases moving from one institution to another, while at risk of recurrence of CDiff, should be fewer still, so that discovery of such cases will represent success with truly rare events.

Conclusion

Along with ten other CDRNs, CAPriCORN is at the halfway point of its "Phase I" life span and is ready to test its systems with real use cases. The infrastructure was designed to allow for evolution in the data model and increasing complexity of queries in the future. Five submitted cohort studies are currently being processed through stages of the CAPriCORN workflow, and a number of new study proposals are being prepared.

Sustainability of the architecture will be demonstrated through a number of additional research studies that had not been considered at the proposal stage. These studies provide a valuable challenge to CAPriCORN's proposal triage, patient-centeredness, and external researcher engagement workflows.

Acknowledgements

The authors are principal informaticians on the CAPriCORN project. Acknowledgement is due to the overall project PI, Terry Mazany of the Chicago Community Trust for his leadership, and to site PIs: other than those listed as co-authors, Fred Rachman of the Alliance, Raj Shah at Rush, and Brian Schmitt at VA Hines; also Jerry Krishnan who was the original site PI at UIC. Doriane Miller and Madeleine Shalowitz direct the PCAC and have provided great clarity on patient engagement. Jonathan Tobin (NYC CDRN) has been a great supporter with ideas and proposals in the External Researcher Committee. Last but not least, John Collins and Shelly Sital at the Illinois Medical District Commission have provided a consistently high standard of project support.

References

[1] Patient Centered Outcomes Research Institute [Internet]. PCOR. [cited 2014 Dec 22]. Available from http://www.pcori.org/content/research-we-support

[2] Abel N. Kho, John P. Cashy, *et al*. Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago. (Under review at JAMIA.)

[3] HealthLNK Data Repository [Internet]. [cited 2014 Dec 22]. Available from: http://www.healthlnk.org/

[4] Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Ann Intern Med 2009; 151:341-4.

[5] PopMedNet [Internet]. [cited 2015 Apr 9]. Available from: http://www.popmednet.org

Address for correspondence

Anthony Solomonides, Center for Biomedical Research Informatics, NorthShore University HealthSystem, 1001 University Place, Evanston, IL 60201, USA. Email: asolomonides@northshore.org

Abel Kho, Division of General Internal Medicine, Northwestern University Feinberg School of Medicine, Rubloff Building 10th Floor, 750 N Lake Shore Drive, Chicago IL 60611, USA. Email: Abel.Kho@nmff.org