Journal of the American Medical Informatics Association, 26(7), 2019, 594–602
doi: 10.1093/jamia/ocz012
Advance Access Publication Date: 2 April 2019
Research and Applications

Research and Applications

Data linkages between patient-powered research networks and health plans: a foundation for collaborative research

Abiy Agiro,[1] Xiaoxue Chen,[1] Biruk Eshete,[1] Rebecca Sutphen,[2] Elizabeth Bourquardez Clark,[2] Cristina M. Burroughs,[2] W. Benjamin Nowell,[3] Jeffrey R. Curtis,[4] Sara Loud,[5] Robert McBurney,[5] Peter A. Merkel,[6] Antoine G. Sreih,[6] Kalen Young,[7] and Kevin Haynes[1]

[1]HealthCore, Wilmington, Delaware, USA, [2]Heath Informatics Institute, University of South Florida, Tampa, Florida, USA, [3]Global Healthy Living Foundation, Upper Nyack, New York, USA, [4]Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, Alabama, USA, [5]Accelerated Cure Project, Waltham, Massachusetts, USA, [6]Division of Rheumatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA and [7]Vasculitis Foundation, Kansas City, Missouri, USA

Corresponding Author: Abiy Agiro, PhD HealthCore, Inc., 123 Justison Street, Suite 200 Wilmington, DE 19801-5134, USA (aagiro@healthcore.com)

Received 17 October 2018; Revised 8 January 2019; Editorial Decision 14 January 2019; Accepted 15 January 2019

ABSTRACT

Objective: Patient-powered research networks (PPRNs) are a valuable source of patient-generated information. Diagnosis code-based algorithms developed by PPRNs can be used to query health plans' claims data to identify patients for research opportunities. Our objective was to implement privacy-preserving record linkage processes between PPRN members' and health plan enrollees' data, compare linked and nonlinked members, and measure disease-specific confirmation rates for specific health conditions.

Materials and Methods: This descriptive study identified overlapping members from 4 PPRN registries and 14 health plans. Our methods for the anonymous linkage of overlapping members used secure Health Insurance Portability and Accountability Act–compliant, 1-way, cryptographic hash functions. Self-reported diagnoses by PPRN members were compared with claims-based computable phenotypes to calculate confirmation rates across varying durations of health plan coverage.

Results: Data for 21 616 PPRN members were hashed. Of these, 4487 (21%) members were linked, regardless of any expected overlap with the health plans. Linked members were more likely to be female and younger than nonlinked members were. Irrespective of duration of enrollment, the confirmation rates for the breast or ovarian cancer, rheumatoid or psoriatic arthritis or psoriasis, multiple sclerosis, or vasculitis PPRNs were 72%, 50%, 75%, and 67%, increasing to 91%, 67%, 93%, and 80%, respectively, for members with $\geq$5 years of continuous health plan enrollment.

Conclusions: This study demonstrated that PPRN membership and health plan data can be successfully linked using privacy-preserving record linkage methodology, and used to confirm self-reported diagnosis. Identifying and confirming self-reported diagnosis of members can expedite patient selection for research opportunities, shorten study recruitment timelines, and optimize costs.

Key words: patient-powered research networks, patient-reported information, anonymous linkage methods, data hashing, claims-based computable phenotypes

INTRODUCTION

Linking digital patient data from diverse sources to health plan administrative claims data can enhance identification for appropriate, targeted treatment options and pinpoint opportunities for patient participation in patient-centered outcomes research (PCOR) such as drug safety and clinical effectiveness research.[1–3] PCOR promotes collaboration and partnership between communities of people with commonly shared health concerns and researchers, patients, clinicians, policymakers,[4–7] and, more recently, payer stakeholders.

In 2013, the PCOR Institute launched its patient-centered data research network, PCORnet.[2,8,9] PCORnet is a distributed network of 13 clinical data research networks, 20 patient-powered research networks (PPRNs), and 2 Health Plan Research Networks.[10] PPRNs are communities of motivated patients and care partners, among others, with common interests in 1 or a group of related diseases, and represent an invaluable source of patient-generated information.[1,2] As part of their mandate, PPRNs include patients in their governance.[11] PPRN members actively reach out and volunteer or are recruited and encouraged to participate in clinical studies.[1,2,12]

To expand their scope of activities and research collaborations, some PPRNs have been engaged with payer stakeholders. Specifically, this study is an example of such engagement—between 4 PPRNs and 14 health plans to explore improvements in methodologies for integrating longitudinal payer claims data into the PCORnet environment. In these engagements, the integration of administrative claims data by deterministic or probabilistic matching based on personal identifying information or potentially utilizing anonymous data linkages is essential. The integrated data environment can be used to validate claims-based computable phenotypes (eg, diagnosis code-based algorithms to identify patients for research opportunities) while using patient-reported disease label from PPRN members as the "silver standard" (the "gold standard" being a review of medical records). It could also be used for evaluating the outcomes of health plan recruitment efforts to increase the size of PPRN membership.

While methods exist to link PPRN data to health plan data, a number of technical challenges, such as incomplete capture of relevant longitudinal data[13] and inadequate data harmonization (standardization), persist.[14] Challenges regarding governance, such as setting policies covering data use and sharing across the PPRN organization[9] and regarding security and privacy of PPRN members, such as requiring sensitive data like social security numbers (SSNs) and technical challenges (eg, implementing privacy-preserving record linkage [PPRL] software solutions)[1,2] continue to impede progress. Other challenges include ethical and regulatory considerations and ongoing monitoring of human subjects' research can be slow, inefficient, and expensive.[15,16]

The main objective of this study was to compare linked and nonlinked members to assess selection bias—on the likelihood of patients joining a PPRN as members and participating in PPRN research—to demonstrate generalizable characteristics and components of already linked PPRN and health plan members. We also aimed to implement a PPRL process between data from 4 disease-specific PPRNs and enrollee membership information from 14 health plans, and measure patient overlap and confirmation rates in specific conditions of interest to the PPRNs. To our knowledge, these analyses will be novel additions to the literature.

MATERIALS AND METHODS

Study design and data sources

This descriptive study used the HealthCore Integrated Research Environment (HIRE) to identify overlapping members between 4 disease-specific PPRNs and 14 geographically dispersed commercial health plans. The HIRE is a repository of longitudinal patient-level administrative claims data for approximately 60 million enrollees and is broadly representative of the United States commercially insured population.[17] This nonexperimental study received Institutional Review Board approval. Researchers accessed a limited, deidentified dataset, and all data were handled in strict compliance with applicable privacy rules under the Health Insurance Portability and Accountability Act (HIPAA).

PPRNs

The study population comprised of members from the 4 disease-specific PPRNs, which are managed by patient-governance groups and are a part of PCORnet.

- The ABOUT (American BRCA Outcomes and Utilization of Testing) Network (aboutnetwork.org) includes men and women 18 years of age and older. Members may have a known genetic mutation (within their family) or a personal or family history of breast, ovarian, or related cancers.[18]
- ArthritisPower concentrates on musculoskeletal and inflammatory skin conditions (focused on arthritis or psoriasis) and operates a nationwide research registry network of patients diagnosed with rheumatoid arthritis, psoriatic arthritis and spondyloarthritis (eg, psoriatic arthritis and ankylosing spondylitis), and a variety of other rheumatic conditions.[19]
- The iConquerMS PPRN specializes in multiple sclerosis (MS). iConquerMS is working to establish a community of 20 000 participants. People with MS and other stakeholders enroll on the network's portal (iConquerMS.org), which facilitates collection of demographic, MS history, and patient-reported outcomes data plus ongoing interactions and communications with the network's members.[20]
- The Vasculitis PPRN (VPPRN) focuses on vasculitis, and has more than 2500 members enrolled in clinical studies investigating multiple types of vasculitis.[21]

Linkage methodology

The anonymous linkage methods we used were built on a secure HIPAA-compliant double-salted SHA (secure hash algorithm)-256 hash function, to conduct PPRL between the HIRE and PPRN databases.[22–24] Hashing was appropriate because the PPRNs and health plans involved elected not to exchange encrypted fully identifiable patient information that could be reversed.

The anonymous linkage of HIRE and PPRN data networks is similar to an approach formulated by Weber et al,[25] which complied with the HIPAA minimum privacy policies, and precluded the full exchange of identifiers such as SSN, which are required by more sophisticated linkage algorithms. As a result, only minimal information was included—patients' whole first and whole last names, dates of birth, and sex, and we avoided using unique distinct identifiers such as SSN. Exact matches on whole first names, whole last names, dates of birth, and sex were used in a deterministic fashion to establish linkage. The study used 2 software implementations of the anonymous linkage algorithm, SQL Server, and JAVA programming languages. Both languages were selected because they were easily implementable in the PPRN data environments.

Linkage scope

Hashing was not limited to PPRN members we expected to compare or link. All patients' data in each of the 4 PPRNs were hashed, irrespective of geography, or whether they reported that they were commercially insured or any other features. The anonymous linkage between the HIRE and each PPRN enabled identification of PPRN network specific overlap in membership, and validation of computable phenotypes. The diagnoses self-reported by PPRN members (the denominator) were compared with claims-based computable phenotypes (the numerator) to generate confirmation rates (percentages) across varying durations of health plan coverage (any, or ≥5 years). Only self-reported diagnosis was used, as not all PPRNs collected other patient information (eg, care from the relevant specialists) or other types of data such as immunosuppressive medication use (eg, biologic therapy).

Computable phenotypes

Both broad and strict confirmation rates for computable phenotypes were calculated for PPRN members who were successfully linked with administrative claims data; specific codes are shown in Supplementary Table 1. Across the 4 PPRNs, broad definition computable phenotypes were based on at least 1 diagnosis in any position that was specific to the condition of interest to the specific PPRN, based on medical claims from any treatment setting including inpatient hospitalization, emergency department services, or outpatient or office visits. Strict definition computable phenotypes relied on more stringent requirements and varied across the cohorts. In the ABOUT PPRN, strict computable phenotypes required at least 2 diagnoses in any position as found in medical claims that were 30 days apart in the office visit setting.[26] Strict computable phenotypes for members of the iConquerMS PPRN required at least 3 claims for MS diagnosis–related hospitalizations or MS diagnosis–related outpatient or emergency department visits in any diagnosis position or MS-related prescription fills in any combination that were no more than 365 days apart.[27] In the ArthritisPower PPRN, strict computable phenotypes required at least 2 diagnoses in medical outpatient claims from a specialist, such as a dermatologist for psoriasis or a rheumatologist for other relevant conditions, and age at diagnosis.[28] Strict computable phenotypes in the VPPRN were constructed from a combination of diagnosis codes, physician specialty (rheumatology, immunology, nephrology, otorhinolaryngology or pulmonary, cardiology or vascular surgery), and the use of immunosuppressive medications.[29]

Statistical analysis

Descriptive statistics were used to establish patient counts and evaluate demographic and clinical characteristics, including age, sex, region, comorbidities, medical and pharmacy utilization, and costs during health insurance coverage periods of January 2006 to July 2017. Differences in demographic and clinical characteristics were compared using *t* test for continuous variables and chi-square test for categorical variables. The confirmation rate is simply a proportion. Therefore, we determined the exact 95% confidence limits for confirmation rates using binomial random variables.

RESULTS

PPRN and HIRE matching

As informed consent to allow data linkages is mandatory for membership in ArthritisPower and VPPRN, all members in these 2 PPRNs were available for linkage. In contrast, 60% and 87% of iConquerMS and ABOUT Network members, respectively, were available for linkage, as membership in these 2 PPRNs is not tied to mandatory informed consent (Table 1). At the time of this analysis, data for 21 616 PPRN members were available to be hashed, including 5665 members from ABOUT Network; 11 343 from ArthritisPower; 2509 from iConquerMS; and 2099 from VPRRN. Of these, 4487 (21%) of the members were linked to the 14 health plans, including 25% (n = 1435) of ABOUT members; 19% (n = 2166) from ArthritisPower; 22% (n = 543) iConquerMS members; and 16% (n = 343) VPPRN members. A total of 3546 (16%) PPRN members were commercially insured and had at least 1 day of medical coverage (Table 1). A total of 684 (3%) of PPRN members overall had at least 5 years of uninterrupted medical insurance enrollment.

Patient characteristics

Compared with the reference group, that is, health plan members who were not linked with PPRNs but who met broad definition computable phenotype (Table 2), PPRN members linked to health plans were younger (mean age 56 ± 16.5 vs 48 ± 11.6 years), more likely to be women (76% vs 92%), and less likely to reside in the North East (23% vs 18%) ($P < .001$). In general, smaller proportions of PPRN members linked to health plans had more comorbid conditions and smaller proportions had more medical and pharmacy utilization vs the reference group. These apparent differences suggest some bias in patient selection when joining PPRNs.

Relative to the reference group, PPRN members who met claims-based broad definition computable phenotype after linkage to health plans were younger (mean age 56 ± 16.5 vs 50 ± 11.2 years), more likely to be women (76% vs 90%), and less likely to reside in the Northeast (23% vs 20%) ($P < .001$. These PPRN members were less likely to have more comorbid conditions and had higher pharmacy utilization vs the reference group. However, they had similar levels of medical utilization compared with the reference group. These apparent differences suggest some bias in research participation as only PPRN members who meet computable phenotype are eligible for collaborative research with health plans.

Broad definition confirmation rates

Confirmation rate for claims-based computable phenotype using patient self-reported diagnosis as the reference standard are shown in Table 3 when no minimum duration of health plan enrollment was required. Irrespective of the duration of coverage, the confirmation rate for breast or ovarian cancer (ABOUT PPRN) was 72% (95% confidence interval [CI], 68%-76%). Confirmation rates increased with 5 years or more of longitudinal health plan coverage: for breast or ovarian cancer, the confirmation rate was 91% (95% CI, 82%-96%). The confirmation rate for breast cancer was 66% (95% CI, 61%-70%), and for ovarian cancer was 68% (95% CI, 55%-79) for any duration. The confirmation rate at 5 years for breast cancer only increased to 90% (95% CI, 81%-96%). For ovarian cancer only, the confirmation rate increased to 100% (95% CI, 72%-100%). In the ArthritisPower PPRN, the confirmation rate for rheumatoid or psoriatic arthritis or psoriasis for patients with any duration of health plan enrollment (ie, no minimum duration of coverage) was 50% (95% CI, 49%-53%). For rheumatoid arthritis, the confirmation rate was 52% (95% CI, 48%-56%). The confirmation rate for psoriatic arthritis was 52% (95% CI, 43%-60%)

Table 1. Patient counts for PPRN-HIRE matching

Steps	Description	All PPRNs	ABOUT Network	ArthritisPower	iConquerMS	VPPRN
1	PPRN memberships	24 131	6513	11 343	4176	2099
2	PPRN memberships obtained and hashed	21 616 (90%)	5665 (87%)	11 343 (100%)	2509 (60%)	2099 (100%)
3	PPRN memberships linked with 14 health plans (ie, final linkage result)	4 487 (21%)	1435 (25%)	2166 (19%)	543 (22%)	343 (16%)
4	Linked PPRN members who were commercially insured including Medicare Advantage (ie, final study sample)	3546 (16%)	1228 (22%)	1600 (14%)	444 (18%)	276 (13%)
5	Linked PPRN members with at least 5 years of uninterrupted insurance coverage (ie, sample size for sensitivity analysis on members)	684 (3%)	187 (3%)	314 (3%)	116 (5%)	67 (3%)

Values are n (%) using step 2 as denominator. HIRE data contained claims from 14 health plans.

ABOUT: American BRCA Outcomes and Utilization of Testing Network; HIRE: HealthCore Integrated Research Environment; PPRN = patient-powered research network; VPPRN = vasculitis patient-powered research network.

and for psoriasis was 47% (95% CI, 38%-55%), as shown in Table 3.

For patients with 5 years or more of health plan enrollment, rheumatoid or psoriatic arthritis or psoriasis were confirmed in claims at 67% (95% CI, 60%-73%). The confirmation rates for rheumatoid or psoriatic arthritis increased to 67% (95% CI, 59%-74%) and 67% (95% CI, 47%-83%), respectively. The confirmation rates for psoriasis also increased to 79% (95% CI, 54%-94%), as shown in Table 4. For MS (iConquerMS), the confirmation rate was 75% (95% CI, 71%-79%), for any duration, and at 5 or more years of health plan enrollment, the rate for MS increased to 93% (95% CI, 87%-97%). The confirmation rate for vasculitis (VPPRN) was 67% (95% CI, 59%-74%), and at 5 years or more of health plan enrollment, the rate increased to 80% (95% CI, 67%-90%).

Strict definition confirmation rates

For any duration of health plan enrollment, 60% (95% CI, 55%-64%) of ABOUT Network members were strictly confirmed using the strict definition in claims. The strict definition confirmation rate for breast cancer in claims was 58% (95% CI, 53%-63%), and for ovarian cancer alone, the strict definition confirmation rate was 63% (95% CI, 50%-75%). Confirmation rates increased with 5 years or more of longitudinal health plan coverage: for breast or ovarian cancer, the strict definition confirmation rate in claims was 90% (95% CI, 80%-96%). Using the strict definition, the confirmation rate for breast cancer alone was 89% (95% CI, 79%-95%), and for ovarian cancer alone, the strict definition confirmation rate in claims was 91% (95% CI, 59%-100%). In the ArthritisPower PPRN, the strict definition confirmation rate for rheumatoid or psoriatic arthritis or psoriasis was 35% (95% CI, 32%-38%) for members with any duration of health plan coverage. For rheumatoid arthritis, the strict definition rate was 37% (95% CI, 33%-41%), and the strict definition confirmation rate for psoriasis was lower, 16% (95% CI, 10%-23%), as shown in Table 3. The strict definition confirmation rate for rheumatoid or psoriatic arthritis or psoriasis for members with 5 years or more of enrollment was 58% (95% CI, 51%-65%). The strict definition confirmation rates of rheumatoid or psoriatic arthritis increased to 59% (95% CI, 51%-67%) and 47% (95% CI, 28%-66%), respectively. The strict definition confirmation rates for psoriasis was 47% (95% CI, 25%-71%), as shown in Table 4. For MS (iConquerMS), the strict definition confirmation rate was 73% (95% CI, 68%-77%) for any duration, and at 5 years the rate for MS increased to 92% (95% CI, 86%-96%). The strict definition confirmation rate for vasculitis (VPPRN) was 42% (95% CI, 35%-49%) for any duration, and at 5 years it increased to 51% (95% CI, 37%-65%).

DISCUSSION

This study examined overlapping membership between 4 disease-specific PPRNs and 14 health plans. From more than 20 000 individual patient records obtained and hashed with PPRL, between 16% and 25% were successfully linked to health plans across the PPRNs. This linkage was performed without the requirement for PPRN members be enrolled in any type of commercial insurance or the 14 health plans studied.

Accessing health plan claims data brings in additional information capable of serving as a foundation for future collaborative research with PPRNs. Claims data capture a comprehensive set of inpatient and outpatient medical encounters that could offer improved verification of self-reported comorbidities by PPRN members. Claims data also allow the assessment of medication adherence, as it captures filled or dispensed prescriptions, which is closer to actual consumption compared with prescription ordering data from electronic medical records. In addition, claims data can be linked to sources of mortality data which could support outcomes research.

Linked patients were about 48 years old, indicative of a working-age population with commercial health plan coverage sponsored by employers. Patients were predominantly women, which was consistent with the conditions focused on by the PPRNs. Hence, the linked members may not be generalizable to broader fee for service Medicare- or Medicaid-insured populations. We also documented how the linked PPRN members differed from nonlinked commercially insured members with conditions of interest. This represents a novel contribution for understanding the selection bias in who is likely to join PPRNs, and who is likely to participate in PPRN research.

Table 2. Summary of PPRN member characteristics

Patient Characteristic	Health Plan Member Not Linked With PPRNs and Met Broad Computable Phenotype (Reference Group) (n = 1 825 115)	PPRN Members Linked With Health Plans (n = 3546)	PPRN Members Linked With Health Plans vs Reference Group (*P* Value)	PPRN Members Linked With Health Plans and Met Broad Computable Phenotype (Numerator) (n = 1293)	Numerator Population of PPRN Confirmation Rate vs Reference Group[a] (*P* Value)
Age, y, mean ± SD	56 ± 16.5	48 ± 11.6	.001	50 ± 11.2	.001
Female	1 386 014 (76)	3256 (92)	.001	1165 (90)	.001
Health plan medical coverage			.001		.001
Commercial	1 536 366 (84)	3383 (95)		1203 (93)	
Medicare Advantage	288 749 (16)	163 (5)		90 (7)	
Duration of medical coverage, y	4 ± 3.4, 3 (1-7)	3 ± 2.9, 2 (1-4)	.001	4 ± 3.3, 3 (1-6)	.001
Urban resident	1 443 470 (79)	2879 (81)	.002	1079 (83)	.001
Census region			.001		.007
Northeast	416 836 (23)	629 (18)		255 (20)	
Midwest	418 593 (23)	874 (25)		327 (25)	
South	537 947 (29)	1123 (32)		368 (28)	
West	445 460 (24)	906 (26)		342 (26)	
Missing	6279 (0.3)	14 (0.4)		1 (0.1)	
Comorbid history					
Chronic pulmonary disease	530 228 (29)	881 (25)	.001	371 (29)	.776
Cardiovascular disease	426 455 (23)	405 (11)	.001	197 (15)	.001
Diabetes	378 262 (21)	437 (12)	.001	185 (14)	.001
Asthma	294 465 (16)	629 (18)	.009	266 (21)	.001
Osteoporosis	345 105 (19)	370 (10)	.001	212 (16)	.021
Medical utilization history					
Any hospitalization	642 938 (35)	937 (26)	.001	418 (32)	.029
Any emergency room visit	789 151 (43)	1273 (36)	.001	536 (41)	.195
Primary care visits	1 566 047 (86)	2727 (77)	.001	1103 (85)	.607
Specialist visits	1 721 255 (94)	2893 (82)	.001	1225 (95)	.503
Lab results available	844 010 (46)	1417 (40)	.001	687 (53)	.001
Pharmacy utilization history					
Members with pharmacy coverage	1 326 550 (73)	2687 (76)	.001	996 (77)	.001
Members with pharmacy fill of those with pharmacy coverage	1 237 639 (93)	2414 (90)	.001	952 (96)	.004

Values are n (%), mean ± SD, or median (interquartile range). Percentages may sum to >100% or <100% due to rounding for census region.

PPRN = patient-powered research network. PPRN members linked with health plans and met broad computable phenotype is also the same as number population of PPRN confirmation rate.

[a]Proportions were compared with chi-square tests while means were compared with *t* tests.

Table 3. Confirmation rates of claims-based diagnosis as percentage of self-reported diagnosis from PPRN members regardless of duration of insurance coverage

			Broad definition confirmation in claims		Strict definition confirmation in claims	
PPRN	Diagnosis from PPRN Member Self-Report	Denominator	n	% (95% CI)	n	% (95% CI)
ABOUT Network	PPRN memberships overlapping with health plans	1228				
	Breast or Ovarian Cancer	519	373	72 (68-76)	309	60 (55-64)
	Breast Cancer	474	311	66 (61-70)	275	58 (53-63)
	Ovarian Cancer	62	42	68 (55-79)	39	63 (50-75)
ArthritisPower	PPRN memberships overlapping with health plans	1600				
	Arthritis or Psoriasis	935	467	50 (49-53)	323	35 (32-38)
	Rheumatoid arthritis	699	357	52 (48-56)	257	37 (33-41)
	Psoriatic arthritis	147	76	52 (43-60)	46	31 (24-40)
	Psoriasis	139	64	47 (38-55)	22	16 (10-23)
	PPRN memberships overlapping with health plans	444				
iConquerMS	Multiple sclerosis	444	335	75 (71-79)	323	73 (68-77)
VPPRN	Number of PPRN membership overlapping with health plans	276				
	Vasculitis	177	118	67 (59-74)	74	42 (35-49)

Confirmation rate for claims-based computable phenotype using patient self-reported diagnosis as the reference standard. No minimum duration of health plan enrollment required.

ABOUT : American BRCA Outcomes and Utilization of Testing; CI: confidence interval; PPRN: patient powered research network; VPPRN = vasculitis patient-powered research network.

Table 4. Confirmation rates of claims-based diagnosis as percentage of self-reported diagnosis for PPRN members with 5 or more years of uninterrupted insurance coverage

		Denominator	Broad Definition Confirmation in Claims		Strict Definition Confirmation in Claims	
PPRNs	Diagnosis From PPRN Member Self-Report	N	n	% (95% CI)	n	% (95% CI)
ABOUT Network	PPRN memberships overlapping with health plans	187				
	Breast or ovarian cancer	78	71	91 (82-96)	70	90 (80-96)
	Breast cancer	72	65	90 (81-96)	64	89 (79-95)
	Ovarian cancer	11	11	100 (72-100)	10	91 (59-100)
ArthritisPower	PPRN memberships overlapping with health plans	314				
	Arthritis or psoriasis	199	133	67 (60-73)	115	58 (51-65)
	Rheumatoid arthritis	160	107	67 (59-74)	94	59 (51-67)
	Psoriatic arthritis	30	20	67 (47-83)	14	47 (28-66)
	Psoriasis	19	15	79 (54-94)	9	47 (25-71)
iConquerMS	PPRN memberships overlapping with health plans	116				
	Multiple sclerosis	116	108	93 (87-97)	107	92 (86-96)
VPPRN	PPRN memberships overlapping with health plans	67				
	Vasculitis	51	41	80 (67-90)	26	51 (37-65)

Confirmation rate for claims-based computable phenotype using patient self-reported diagnosis as the reference standard. Five or more years of health plan enrollment was required.

ABOUT: American BRCA Outcomes and Utilization of Testing; PPRN: patient powered research network; VPPRN = vasculitis patient-powered research network.

We demonstrated that it was possible to securely link and confirm patient generated data from PPRN and health plans. The open-source, privacy-preserving linkage processes we used represent scalable, low-cost options for other PPRNs and registries, and are devoid of key restrictions such as end-user licensing and other costly impositions. Restrictions like end-user licensing, even for open-source programs, may require time and resources to draft and implement agreements. This study took advantage of the scalability of this privacy preserving approach because no exchange of sensitive data fields such as SSN were required, nor was there any exchange of protected health information. In addition, scalability was facilitated by the freedom to select from different programming languages including the more established SQL as well as newer software products such as Java.

Considering PPRN membership as a veritable "silver standard" disease label compared with claims data, our findings support the concept that simpler and broader phenotype definitions may be just as good as stricter ones (the "gold standard" being a review of medical records). This study also demonstrated that claims-derived confirmation rates increased in direct proportion to the duration of

health plan enrollment. Improved confirmation of PPRN conditions was found in patients with longer uninterrupted insurance coverage with a health plan. Therefore, linking of PPRN membership to health plan data may yield greater value if they link to health plan databases with greater local market share and longer longitudinal coverage.

Nonetheless, the process of linking data across different networks is fraught with challenges. Governance,[9] security, and privacy issues, while exceptionally challenging now, may become even more difficult with future cyber threats.[1,2] Lapses in any of these areas could seriously interfere with patient trust, which is essential for PCOR initiatives. To understand and address the data linkage trust issue on the patient side, we are conducting in-depth interviews of patient participants from PPRNs. In addition, we intend to produce an educational video on PPRL process and disseminate our result back to PPRN members.

Currently, health plan enrollees do have access to their own claims data through health plan member portals. Similarly, participants of PPRNs get access to their own data contributed to the PPRN. Linking together PPRN data with health plan claims helps to build the foundation for future research. The HIRE repository affords researchers an important strength such as large local market penetration and longitudinal follow-up that may not be available in other large health plan data repositories. While the researchers in this study only accessed a limited set of de-identified data, information curated in the HIRE allow for full patient identification and may be used to communicate with and recruit patients for studies, and is ideally suited for pragmatic clinical trials. Furthermore, this approach to patient identification, particularly strict computable phenotypes, could be made substantially more stringent depending on the need, and be invaluable in patient recruitment in future studies. Toward this end, we have identified and randomized health plan members who are not currently part of the 4 PPRNs into mail vs email recruitment groups, after identifying them through the now confirmed strict definition computable phenotypes. The findings of that effort will soon be reported through a separate publication.

Limitations

Patients in HIRE and the PPRNs were identified with diagnosis and procedure codes. Administrative claims may have coding inaccuracies resulting in outcome misclassifications, and over- or underestimation of the sample. Claims may have incomplete clinical data capture, which interfere with estimates. Deterministic matching is overly conservative and may not link in situations with spelling or naming differences (eg, "Robert" in HIRE data, but "Bob" in the PPRN data). Deterministic matching is not as flexible as more sophisticated and probabilistic matching algorithms, which come at the cost of reduced trust by patients, as they require disclosure and full exchange of identifiers, either multiple nonunique identifiers[30] or sensitive identifiers including SSN. As PPRNs rely on patient trust for sustainability, deterministic matching, through PPRL, is a reasonable approach.

This process did not have internal validation through manual or other review of matches using fully identifiable data exchange as that was not possible in a PPRL approach, which is not a reversible process. It will be instructive for future research efforts to explore fully identifiable record linkage and compare the results to PPRL. Such research, however, will be challenging given the legal, governance, data security, and trust hurdles including the possible need to obtain informed consents from all members of health plans and PPRNs alike. Currently, health plans do not have the legal authorization to require enrollees to consent to fully identifiable data linkage. Further, it is doubtful that an institutional review board or any privacy board can have jurisdiction to authorize waiver of informed consent for the transfer of fully identifiable information of 60 million health plan members to PPRNs for internal validation of data linkages.

All enrollee data in the HIRE and hence the linked PPRN members have commercial health insurance coverage including Medicare Advantage members. As a result, enrollment rates were affected by turnover for a variety of reasons including members switching health plans. In addition, because the study populations was commercially insured, these results may not be readily generalizable to Medicaid, which may have differing levels of access to healthcare resources, insufficient educational preparation to understand and participate in the offerings of PPRN-type initiatives, or are precluded from participation because of logistic or geographic considerations. The challenges of obtaining state-by-state government permissions to include Medicaid data for research are not limited to PCORnet, and are also a feature of other similar efforts such as Sentinel.[31]

Data were hashed, and linkage was attempted, regardless of what type of insurance the PPRN member reported. It is probable that restricting the linkable sample to patients who self-reported coverage from 1 of the 14 health plans studied would have greatly improved the linkage rates beyond the 16%-25% we observed. Finally, while some self-reported PPRN conditions like breast cancer and vasculitis are highly specific, the confirmation rate of self-reported rheumatoid arthritis, for example, in the absence of additional data is relatively low.[32] For this reason, requiring additional information from patients to improve the specificity of self-reported conditions (eg, for rheumatoid or psoriatic arthritis, active care from a rheumatologist, plus use of disease modifying antirheumatic drugs or biologic therapy) increases the specificity of the condition and likely would have increased the confirmation rates beyond the 67% we observed in claims data for this study.

CONCLUSION

This study demonstrated the ability to successfully link health plan patient-level data and PPRN membership data while safeguarding privacy and security with the judicious use of anonymous linkage. The selection bias on who joins a PPRN as a member is similar to the bias on who ultimately joins a PPRN research opportunity. Higher claims-based confirmation rates of PPRN conditions were found for self-reported diagnosis from patients with longer uninterrupted insurance coverage in the same health plan. The strict definition computable phenotypes analyzed in this study can be used to identify health plan members who are not currently members of PPRNs to be invited to join PPRNs and engage in related research opportunities, while potentially shortening study recruitment timelines and reducing research costs.

FUNDING

This work was supported by Patient-Centered Outcomes Research Institute awards covered under "A model for improving patient engagement and data integration with PCORnet Patient Powered Research Networks and payer stakeholders" (ME-1503-28785). ArthritisPower was supported through PCOR Institute (PCORI) award PPRN-1306-04811. ABOUT (American BRCA Outcomes and Utilization of Testing) was supported through PCORI

award PPRN-1306-04846. iConquerMS (Multiple Sclerosis) was supported through PCORI award PPRN-1306-04704. VPPRN (Vasculitis Patient-Powered Research Network) was supported through PCORI award PPRN-1306-04758.

AUTHOR CONTRIBUTIONS

KH, JRC, PAM, RS, RM, SL, and KY made substantial contributions to the conceptualization and design of this research project as well as manuscript planning. AA, XC, BE, WBN, CB, EBC, and AGS contributed substantially to the acquisition, analysis, and interpretation of the data for this research. AA and XC were responsible for organizing and drafting the manuscript content, while KH, JRC, PAM, RS, RM, SL, KL, BE, WBN, CB, EBC, and AGS contributed critical revisions and reviewed to ensure accuracy and applicability of the manuscript's intellectual content. All the authors take responsibility for the final approval of the version to be published, and are accountable for all aspects of the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

Bernard B. Tulsi, senior medical writer at HealthCore, provided writing and editorial support for this article. Lauren Parlett and Jeffrey Greenberg, senior researchers at HealthCore, have reviewed the initial draft and provided valuable comments. Jessee Young and David (Marc) Cram, senior developers at HealthCore, have created and tested the hashing algorithm. We thank our HealthCore (Dianna Hayden, Zhengzheng Jiang, Michael Mack), iConquerMS (William Tulskie, Leonid Kagan, Kenneth Buetow), and Arthritis-Power (Lang Chen, Shou Yang, Robert Matthews) colleagues for programming support. We also thank Sue Friedman and Marleah Dean Kruzel, patient representatives from the ABOUT PPRN's Executive Committee, for critiquing the draft to enhance this article. We also thank Kelly Gavigan (data scientist) and Kelly V. Clayton (patient representative) from Global Healthy Living Foundation for administrative support.

All statements in this manuscript, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee or other participants in PCORnet.

CONFLICT OF INTEREST STATEMENT

RS received payment for serving as Chief Medical Officer of InformedDNA, which provides a network of genetics specialists and genetics benefit management services in partnership with health plans, including Anthem, Inc. KH, AA, LC, BE, KY, WBN, AGS, SL, EBC, CB, PAM, JRC, and RM report no conflicts.

REFERENCES

1. Fleurence RL, Beal AC, Sheridan SE, *et al*. Patient-powered research networks aim to improve patient care and health research. *Health Aff (Millwood)* 2014; 33 (7): 1212–9.
2. Fleurence RL, Curtis LH, Califf RM, *et al*. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; 21 (4): 578–82.
3. West SL, Johnson W, Visscher W, *et al*. The challenges of linking health insurer claims with electronic medical records. *Health Informatics J* 2014; 20 (1): 22–34.
4. Forsythe LP, Ellis LE, Edmundson L, *et al*. Patient and stakeholder engagement in the PCORI pilot projects: description and lessons learned. *J Gen Intern Med* 2016; 31 (1): 13–21.
5. Frank L, Basch E, Selby JV; Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. *JAMA* 2014; 312 (15): 1513–4.
6. Selby JV. Linking automated databases for research in managed care settings. *Ann Intern Med* 1997; 127 (8 Pt 2): 719–24.
7. Selby JV, Lipstein SH. PCORI at 3 years–progress, lessons, and plans. *N Engl J Med* 2014; 370 (7): 592–5.
8. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014; 21 (4): 576–7.
9. Consortium PCP, Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. *J Am Med Inform Assoc* 2014; 21: 583–6.
10. PCORNET. The National Petient-Centered Clinical Research Network. https://pcornet.org/participating-networks/ Accessed December 10, 2018.
11. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building data infrastructure to evaluate and improve quality: PCORnet. *J Oncol Pract* 2015; 11 (3): 204–6.
12. Mazor KM, Richards A, Gallagher M, *et al*. Stakeholders' views on data sharing in multicenter studies. *J Comp Eff Res* 2017; 6 (6): 537–47.
13. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013; 51: S30–7.
14. Ogunyemi OI, Meeker D, Kim HE, *et al*. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. *Med Care* 2013; 51: S45–52.
15. Sung NS, Crowley WF Jr, Genel M, *et al*. Central challenges facing the national clinical research enterprise. *JAMA* 2003; 289 (10): 1278–87.
16. Trinidad SB, Fullerton SM, Ludman EJ, *et al*. Research ethics. Research practice and participant preferences: the growing gulf. *Science* 2011; 331 (6015): 287–8.
17. Wasser T, Wu B, Ycas J, Tunceli O. Applying weighting methodologies to a commercial database to project US Census demographic data. *Am J Account Care* 2015; 33–8. http://www.ajmc.com/journals/ajac/2015/2015-vol3-n3/applying-weighting-methodologies-to-a-commercial-database-to-project-us-census-demographic-data/p-2 Accessed June 20, 2017.
18. ABOUT Patient-Powered Research Network (ABOUT Network). http://pcornet.org/patient-powered-research-networks/pprn8-university-of-south-florida/ Accessed June 25, 2017.
19. AR-PoWER (ARthritis Partnership with Comparative Effectiveness Researchers) PPRN. https://www.pcori.org/research-results/2015/ar-power-arthritis-partnership-comparative-effectiveness-researchers-pprn Accessed June 23, 2018.
20. The Multiple Sclerosis Patient-Powered Research Network, iConquerMS. https://www.pcori.org/research-results/2015/multiple-sclerosis-patient-powered-research-network-iconquerms%E2%84%A2 Accessed June 24, 2018.
21. VPPRN. https://www.rarediseasesnetwork.org/cms/vcrc/Research/VPPRN Accessed June 27, 2018.
22. Information Technology Laboratory, National Institute of Standards and Technology. Secure Hash Standard (SHS). Federal Information Processing Standards Publication. United States Department of Commerce; 2012.
23. Kijsanayotin B, Speedie SM, Connelly DP. Linking patients' records across organizations while maintaining anonymity. AMIA Annu Symp Proc 2007, 2007: 1008. Chicago, Illinois.
24. Mushlin AB, Bell C, Brown B, *et al*. *Anonymous Linking of Distributed Databases.* Mini-Sentinel Activities. 2013. US Food and Drug Administration. https://www.sentinelinitiative.org/sites/default/files/data/complementary-data/Mini-Sentinel_Anonymous-Linking-of-Distributed-Databases_0.pdf. Accessed April 17, 2015.
25. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc* 2012; 19 (e1): e157–61.
26. Whyte JL, Engel-Nitz NM, Teitelbaum A, *et al*. An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. *Med Care* 2015; 53 (7): e49–57.
27. Wallin MT, Culpeppe WJ, Campbell J, *et al*. The prevalence of multiple sclerosis in the United States: a population-based healthcare database approach.

ECTRIMS Online Library 2018; 199999. Available at:https://onlinelibrary.ectrims-congress.eu/ectrims/2017/ACTRIMS-ECTRIMS2017/199999/mitchell.t.wallin.the.prevalence.of.multiple.sclerosis.in.the.united.states.a.html

28. Kim SY, Servi A, Polinski JM, *et al*. Validation of rheumatoid arthritis diagnoses in health care utilization data. *Arthritis Res Ther* 2011; 13 (1): R32.
29. Sreih AG, Annapureddy N, Springer J, *et al*. Development and validation of case-finding algorithms for the identification of patients with anti-neutrophil cytoplasmic antibody-associated vasculitis in large healthcare administrative databases. *Pharmacoepidemiol Drug Saf* 2016; 25 (12): 1368–74.
30. Curtis JR, Chen L, Bharat A, *et al*. Linkage of a de-identified United States rheumatoid arthritis registry with administrative data to facilitate comparative effectiveness research. *Arthritis Care Res (Hoboken)* 2014; 66 (12): 1790–8.
31. Haynes K, Lin ND, Avillach P, *et al*. Extending comparative effectiveness research and medical product safety surveillance capability through linkage of administrative claims data with electronic health records: a Sentinel-PCORnet collaboration. https://www.sentinelinitiative.org/sites/default/files/data/complementary-data/Sentinel_Sentinel-PCORnet-White-Paper_0.pdf Accessed October 10, 2018.
32. Mikuls TR, Saag KG, Criswell LA, *et al*. Mortality risk associated with rheumatoid arthritis in a prospective cohort of older women: results from the Iowa Women's Health Study. *Ann Rheum Dis* 2002; 61 (11): 994–9.