EXTENDING COMPARATIVE EFFECTIVENESS RESEARCH AND MEDICAL PRODUCT SAFETY SURVEILLANCE CAPABILITY THROUGH LINKAGE OF ADMINISTRATIVE CLAIMS DATA WITH ELECTRONIC HEALTH RECORDS: A SENTINEL-PCORnet COLLABORATION

Prepared by: Kevin Haynes, PharmD, MSCE,[1] Nancy D. Lin, ScD,[2] Paul Avillach, MD, PhD,[3,4] Thomas W. Carton, PhD, MS,[5] Jeffrey R Curtis, MD, MS, MPH,[6] Kevin Fahey, MA,[7] Crystal Garcia, MPH,[8] Thomas Harkins, MA, MPH,[9] Wenke Hwang, PhD,[10] Cheryl N. McMahill-Walraven, MSW, PhD,[11] David Meltzer, MD, PhD,[12] Eliel Oliveira, MBA, MS,[5] Pamala A. Pawloski, PharmD,[13] Micah Prochaska, MD,[12] Jon Puro, MPA:HA,[14] Nandini Selvam, PhD, MPH,[1] Richard Platt, MD, MSc[8]

Author Affiliations: 1. HealthCore, Inc., Wilmington, DE 2. Optum, Waltham, MA 3. Department of Biomedical Informatics, Harvard Medical School, Boston, MA 4. Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA 5. Health Services Research, Louisiana Public Health Institute, New Orleans, LA 6. Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL 7. America's Health Insurance Plans, Washington, DC 8. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 9. Humana – Comprehensive Health Insights, Louisville, KY 10. Pennsylvania State University College of Medicine, Hershey, PA 11. Data Science, Aetna, Blue Bell, PA 12. Section of Hospital Medicine, University of Chicago, Chicago, IL 13. HealthPartners Institute for Education and Research, Minneapolis, MN 14. OCHIN, Inc., Portland, OR

Acknowledgments: The authors acknowledge substantial input to this white paper by FDA and PCORI staff, Workgroup members, and members of PCORnet networks.

November 24, 2015

Sentinel is an active surveillance system sponsored by the U.S. Food and Drug Administration (FDA) for monitoring the safety of FDA-regulated medical products. Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

This white paper was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (P122013-499A) for development of the National Patient-Centered Clinical Research Network, known as PCORnet. PCORI is an independent, nonprofit organization authorized by Congress in 2010. Its mission is to fund research that will provide patients, their caregivers, and clinicians with the evidence-based information needed to make better-informed healthcare decisions. PCORI is committed to continually seeking input from a broad range of stakeholders to guide its work. PCORnet is an innovative initiative of PCORI. The goal of PCORnet is to improve the nation's capacity to conduct comparative effectiveness research efficiently by creating a large, highly representative network for conducting clinical outcomes research. *The statements presented in this paper are those of the author(s) and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.*

EXTENDING COMPARATIVE EFFECTIVENESS RESEARCH AND MEDICAL PRODUCT SAFETY SURVEILLANCE CAPABILITY THROUGH LINKAGE OF ADMINISTRATIVE CLAIMS DATA WITH ELECTRONIC HEALTH RECORDS: A SENTINEL-PCORnet COLLABORATION

Table of Contents

I. EXECUTIVE SUMMARY 1

II. INTRODUCTION 1

III. CONCEPTUAL FRAMEWORK 2

IV. GOVERNING PRINCIPLES FOR DATA LINKAGE COLLABORATIONS 4

A. ENGAGEMENT OF PARTNERS AS SCIENTIFIC COLLABORATORS 4
B. MINIMUM DATA NECESSARY 4
C. USE OF EXISTING PCORNET AND SENTINEL COMMON DATA MODELS 5
D. DATA SHARING OR BUSINESS ASSOCIATE AGREEMENTS 5

V. TECHNICAL CONSIDERATIONS FOR DATA LINKAGE COLLABORATIONS 5

A. RETRIEVAL OF PERSONALLY IDENTIFIABLE INFORMATION (PII) 5
B. METHODS FOR DATA LINKAGE 6
C. OTHER SOURCES FOR INTEGRATED CLAIMS AND EHR DATA RESOURCES FOR RESEARCH 6

VI. ORGANIZATIONAL-LEVEL OVERLAP 7

A. CDRN ASSESSMENT: OVERLAP BASED ON CDRN IDENTIFIED INSURANCE STATUS USE CASE 7
B. SENTINEL ADMINISTRATIVE HEALTH PLAN DATA PARTNER ASSESSMENT: OVERLAP ASSESSED VIA INSTITUTIONAL PROVIDER IDENTIFIERS USE CASE 8
C. ASSESSMENT OF ORGANIZATIONAL OVERLAP THROUGH USE OF HASHED IDENTIFIERS 8

VII. DATA LINKAGE USE CASES 8

A. TYPOLOGY OF PATIENT-LEVEL LINKAGE COLLABORATIONS 9
B. USE CASES 9
1. Use Case 1: Patient Provides Consent 10
a. Institutional Review Board approval 10
b. Patient population available for data linkage 10
c. Content of data provided 11
d. Data linkage and data flow requirements 11
2. Use Case 2: Waiver of Consent and HIPAA Authorization 11
a. Institutional Review Board approval 11
b. Patient population available for data linkage 12
c. Content of data provided 12
d. Data linkage and data flow requirements 12
3. Use Case 3: Public Health Activity 12
a. Institutional Review Board approval 13

b. Patient population available for data linkage 13
c. Content of data provided 13
d. Data linkage and data flow requirements 13
4. Use Case 4: De-identified Linked Dataset for Multiple Purposes 13

VIII. DISCUSSION 14

IX. REFERENCES 17

X. APPENDIX: EXAMPLE PROCESS FOR TECHNICAL TRANSFER OF DATA 18

A. EXAMPLES 18
1. ADAPTABLE Data Request from CDRNs 18
2. Dabigatran Data Request from Sentinel 19

XI. TABLES & FIGURES 21

A. FIGURE 1. CONCEPTUAL MODEL 21
B. FIGURE 2. WORKING EXAMPLE 22
C. FIGURE 3. SPECTRUM OF COLLABORATION 22
D. TABLE 1. TYPOLOGY OF PATIENT-LEVEL LINKAGE STUDIES 23
E. USE CASE 1 24
F. USE CASE 2 25
G. USE CASE 3 25
H. USE CASE 4 26

I. EXECUTIVE SUMMARY

Sentinel and the Patient-Centered Outcomes Research Network (PCORnet) organizations hold highly complementary data across diverse networks of data partners. Linkage of these data could improve Sentinel's medical product safety surveillance capabilities and enhance PCORnet's conduct of patient-centered comparative effectiveness research (PCOR). A working group composed of Sentinel and PCORnet data partners was formed to develop a conceptual framework and governing principles for bidirectional data linkage collaborations. The workgroup identified a set of organizational and patient-level use cases that would leverage linked information for PCORnet and Sentinel activities, to guide discussion on specific governance issues and technical requirements for performing data linkage. We provide a framework for a stepped approach that allows identification of the size of a shared population, without disclosure of any patient-level data, and offer a consideration of the creation of a persistent linked data resource. Additionally, we provide a use case with patient authorization in the setting of a consented randomized controlled trial and a use case in the setting of an observational research study that relies on a waiver of patient informed consent and HIPAA authorization. Finally, we address the use case of creating a linked dataset to support an array of evaluations. There are compelling motivations to develop governance, linkage policies, procedures, and infrastructure to allow use of the complementary data sources in claims and EHR data for medical product safety surveillance activities and patient-centered comparative effectiveness research. We propose an incremental approach to addressing the technical and governance challenges in data linkage infrastructure across distributed data networks.

II. INTRODUCTION

The US Food and Drug Administration (FDA) Sentinel System and the Patient-Centered Outcomes Research Institute's (PCORI) PCORnet have developed unique, valuable, and highly complementary distributed data systems containing information from tens of millions of individuals.[1] The FDA Sentinel System monitors the safety of FDA regulated medical products using electronic healthcare data arising principally from large datasets of administrative claims maintained by insurance companies.[2] PCORnet, focused on both observational and interventional comparative effectiveness research (CER), uses electronic health record (EHR) data from clinical data research networks (CDRNs), together with patient-generated data from patient-powered research networks (PPRNs), as its foundation.[3] PCORnet CDRNs are built on a foundation of EHR data, providing electronic health information recorded during routine patient care. PCORnet PPRNs are registries of patients for the study of specific disease states of interest with the ability to collect patient-reported outcomes in large numbers of patients.[4]

Each program's data has strengths and limitations. The Sentinel System's data partners have a complete longitudinal record of reimbursed medical care during clearly defined periods in which an individual has insurance coverage. Insurance claims provide substantial information about diagnoses, procedures (including treatments like immunizations), and outpatient pharmacy dispensing for individuals with medical coverage and pharmacy benefits. There can be reasonable confidence that if no claim exists, especially for a serious condition such as an acute myocardial infarction, the individual did not receive care for this condition during the period of insurance coverage. Similar to EHR data systems, over-the-

counter medication utilization and care that does not arise to medical attention is not recorded. Some of the Sentinel data partners, such as the Kaiser Permanente partners, have access to integrated clinical data, but the largest share of membership in Sentinel comes from the large commercial insurance health plan data partners.[2] The Sentinel large administrative health plan data partners have member populations with representation across the entire US but they either lack access or, in the case of Optum, have variable access to integrated de-identified claims and EHR data for some of their insured populations.

PCORnet's strength is the substantial clinical detail available in EHRs, including both structured elements like vital signs and detailed narratives, plus patient-generated data. EHR based systems are also the only source of data, albeit of variable completeness, for individuals without insurance coverage. The PCORnet EHR data sources are typically restricted to specific institutions, so it is not possible to observe care received in other facilities.

Given the complementary nature of the health plan enrollment- and claims-based longitudinal information collected by Sentinel administrative health plan data partners and the detailed clinical and patient-generated information captured by the PCORnet data partners, it would be highly desirable to link these data for the several million people who are represented in both systems.[5] Establishing such linkages would combine the benefits of administrative claims, EHR, and patient-generated data, enhancing the ability of both PCORnet and the Sentinel System to better achieve their separate aims. The goal in distributed research is to maintain data in source systems as the preferred approach, however, sometimes a minimum necessary person-level data linkage is required to address a specific aim. Quality improvement, while important both within and across health systems is beyond the scope of our presentation. Our objectives in this white paper are to provide a conceptual framework to facilitate data linkage discussions by identifying governing principles for the development of data linkage collaborations, and highlight additional technical and governance requirements via a discussion of selected use cases to achieve the goal of developing bi-directional data collaborations between Sentinel and PCORnet data partners to support the conduct of comparative effectiveness research and medical product safety surveillance.

While there is clear benefit for both PCORnet and Sentinel to establish a readily accessible, up to date, and linked dataset, the governance challenges of determining where a persistent linked dataset would reside, how often to update the data resource, and who has access to the resource are formidable. We thus include in this discussion consideration of a master patient record locator (MPRL) for rapid public health and study feasibility assessment. The focus of the current discussion is on building a conceptual framework and governing principles while highlighting use cases at the organizational and patient-level using specific proposed research studies.

III. CONCEPTUAL FRAMEWORK

Figure 1 is a conceptual framework of a patient interacting with multiple healthcare delivery systems. Each entity that delivers healthcare has unique privacy and consent issues and concerns about the use of data related to their practice patterns and outcomes. Each line represents opportunities and challenges in data governance; legal, regulatory, and compliance concerns; and technical approaches in data transformation. Patients are at the center of healthcare delivery and seek care from multiple

organizations for clinical care. Patients and care delivery organizations (e.g. health systems affiliated with CDRNs) also interact with insurance providers (payers) to provide reimbursement for the care delivered across the healthcare system. In addition, patients can contribute patient-generated data, such as patient-reported outcomes (PROs), to PPRN registries or other entities that may have further interactions with delivery systems or payers. Each use case presented below addresses the technical and data governance concerns of patients and healthcare entities and the relationships between health systems and health plans to form data linkage collaborations for medical product safety surveillance and comparative effectiveness research. Governing principles for protection of patient privacy will also be presented. Additionally, both payers and providers maintain sensitive data for the operation of their organizations that require additional safeguards and strict data use agreements to provide data governance. Finally, PPRNs and registries often contain additional sensitive data on specific groups of patients that must be considered.

Figure 2 presents a simplified working example of a patient engaged in the healthcare system. The patient has a number of hospital episodes at different providers over a defined period with multiple provider encounters, including two providers who contribute clinical EHR data to datamarts within a CDRN participating in PCORnet. A CDRN may be composed of multiple health systems that each have their own datamart. For example hospital A and affiliated clinics may be on one side of town and hospital B and affiliated clinics may be on the other side of town. Additionally, the patient has clinical encounters at laboratories, pharmacies, and receives services outside of both hospital health systems participating in PCORnet, e.g., encounters with providers not affiliated with either hospital system. Over the same period, this patient is enrolled in one health insurance plan, who is a Sentinel administrative health plan data partner. In this example, the healthcare claims associated with all reimbursed services (i.e., providers participating in PCORnet and those not participating in PCORnet) throughout this period would be captured by the Sentinel administrative claims data partner, while some of the detailed clinical EHR information associated with specific encounters would be captured by the providers participating in PCORnet. This example highlights many of the complicated data relationships exhibited in the conceptual model presented. The use cases below will draw on this example to present the proposed framework for CER and medical product safety surveillance enhanced through bidirectional data linkages.

The ideal setting for the conduct of CER and medical product safety surveillance is administrative claims data combined with EHR and clinic-based and/or patient-based registry data (e.g. from a PPRN). This ideal setting would create a persistent linked dataset with controlled access through a trusted third party. The multiuse dataset would allow for rapid medical product safety surveillance activities and CER study feasibility assessment. There is considerable precedent for the sort of linked dataset that could fulfill this ideal setting in the Accountable Care Organization data models in which payers routinely provide complete claims data for the purposes of clinical care coordination, however institutional guidelines for sharing personal health information vary considerably, and may be more restrictive for purposes other than treatment, payment, or operations. Addressing the governance challenges to facilitate a similar shared data model for research would appeal to both the medical product safety surveillance activities of Sentinel and the PCOR activities of PCORnet.

Short of the ideal setting, our goal is to identify opportunities and approaches to readily create linked EHR and claims data resources on an as-needed basis to address specific CER projects or medical product safety surveillance questions. An essential first step is to assess the potential overlap in

membership between populations. There exists a spectrum of potential collaboration, ranging from identifying the extent of overlap at an organizational level to the needs of identifying specific patient-level overlap for prospective longitudinal follow-up studies (**Figure 3**). We present specific use cases below to highlight specific opportunities and approaches to link EHR and claims data resources.

As an intermediate step between a study specific linkage and the creation of a general purpose analyzable linked dataset, the creation of a many-to-many linked dataset of identifiers that indicates the databases a given individual has contributed data to would be an appealing resource for both medical product safety surveillance and rapid feasibility assessment in preparation to observational or pragmatic research. This would entail the creation of either a master patient record locator or additionally through the utilization of anonymous hashed identification with a common hash platform and secure transmission of the salt or seed necessary to generate the hashed identifiers. A common challenge is in ensuring the privacy and confidentiality of the individual patient and we present examples of methods to mitigate these risks, e.g. by introduction of honest brokers and encryption of PHI, using a non-public key (also called a salt or seed) to create the encrypted, or hashed, PHI. The utility of a master patient record locator or hashed identification is appealing to medical product safety surveillance, comparative effectiveness research, and clinical operations. We present use cases below to highlight the potential implementation of a persistent centrally linked EHR and claims data resource, or linkage in a distributed fashion, implemented at some CDRNs.

IV. GOVERNING PRINCIPLES FOR DATA LINKAGE COLLABORATIONS

A. ENGAGEMENT OF PARTNERS AS SCIENTIFIC COLLABORATORS

Linkage of information between data partners requires agreement on governance, collaborative, and technical issues to ensure data security, patient privacy, compliance, and proprietary and collaborative needs. The workgroup identified the need for early and ongoing engagement of partners as scientific collaborators to leverage partner-specific expertise critical to study planning, including input on topics such as characteristics of data, assessments of data quality, and appropriate use of the available data. In the desire to improve efficiencies in surveillance and research activities, it will become necessary to build trust across local organizations in generating persistent standing resources for rapid analytic capabilities. However, data partners from both the Sentinel and PCORnet perspective remain the experts in the data collection and transformation activities. Additionally, local data partner engagement is critical to understanding the local and institutional guidelines for sharing protected health information (PHI) and personally identifiable information (PII) for research, as there may be strict regulations governing data use. If broad linkage to create a complete linked dataset is to be realized, substantial local engagement will be required to address organizational concerns and needs, and to build the bidirectional trust necessary to allow a resource to serve multiple objectives, including medical product surveillance and comparative effectiveness research.

B. MINIMUM DATA NECESSARY

The workgroup agreed that the content of the data shared should be the minimum data necessary to address the specific aims of the projects. Although, from a sustainability perspective larger multiuse datasets such as persistent linked resources could offer efficiencies to support rapid cycle analytics.

C. USE OF EXISTING PCORnet AND SENTINEL COMMON DATA MODELS

Both Sentinel and PCORnet have made substantial investments in development and expansion of CDM data infrastructure and associated analytic tools to support rapid querying capability. Given the extensive data cleaning and curation efforts required to support research and surveillance activities, piggybacking on existing infrastructure where possible and appropriate is preferred, and ultimately more sustainable.

D. DATA SHARING OR BUSINESS ASSOCIATE AGREEMENTS

Data sharing agreements and/or business associate agreements between individual CDRN/PPRN-Sentinel health plan pairs, and between an individual partner and a Coordinating Center, will typically be required. Even with patient authorization (through signed consent) or through a waiver of informed consent granted by an IRB, a specific data sharing agreement with regards to the specified use case of either a single study with consent, an observational longitudinal analysis, or a medical product safety surveillance activity would need to document the allowable uses of PHI and PII. The goal would be to develop a broad agreement to cover multiple future projects (e.g., via creation of persistent linked datasets or a standing infrastructure to query and link datasets on-demand for specific projects with appropriate approvals for rapid analytic capacity). This would require a stepped approach to build bi-directional trust upon prior experience and engagement opportunities.

Separately, these agreements could be extended to include agreements between CDRN to CDRN data partners and between different Sentinel administrative health plan data partners to extend person-time follow-up through the exchange of information longitudinally. This topic is beyond the scope of the present discussion that is focused on enhancing the depth and breadth of data available during a defined period for each individual.

These general principles will guide the discussion of organizational overlap and patient-level overlap presented below. Health systems and health plans are already engaged in data linkage activities for the purposes of treatment, payment, and operations. The health systems that comprise the CDRNs routinely submit claims with identifiable information to health plans to seek reimbursement. Additionally, the exchange of healthcare data is routine, particularly in the setting of accountable care organizations and established health information exchanges. Establishing the use of these data linkage activities for the purposes of medical product safety surveillance or comparative effectiveness research remains challenging as these activities remain beyond the treatment, payment and operation purposes.

V. TECHNICAL CONSIDERATIONS FOR DATA LINKAGE COLLABORATIONS

A. RETRIEVAL OF PERSONALLY IDENTIFIABLE INFORMATION (PII)

The PCORnet and Sentinel common data models (CDMs) include an arbitrary, non-informative patient identifier (e.g, 000001, 000002, …) that is used to follow a patient over time within each data partner's specific data. The omission of medical record numbers, Social Security numbers, and other directly identifiable information is intentional to avoid the possibility of unintended disclosure. Data partners will need to access internal data sources to extract the required PII, and following extraction of the PII,

map the PII to specific patients within the CDMs in order to enable PII patient-level linkage of PCORnet data to Sentinel administrative health plan data partner claims data. Some efficiencies could be gained through expansion of the PCORnet and Sentinel CDMs to include the unique identifiers of both providers and patients needed to conduct routine linkage. Data partners extract, transform, and load data from source systems into the CDMs which are retained behind organizational firewalls and made available for querying. In this process an arbitrary, non-informative patient identifier is generated for unique patients. The CDM could be expanded to include direct identifiers such as names, addresses, and insurance identifiers. This would obviate the need to create internal persistent linkages to PII stored in other information systems outside of a data partner's CDM. However, this expansion could introduce privacy concerns, such as PII being stored in data systems used for routine querying, and would also require substantial effort by both the PCORnet and Sentinel data partners to add new data elements such as direct identifiers to respective CDM data environments. One way to mitigate the privacy concerns would be to store hashed identifiers in the CDM, which would require the technical capacity to maintain a common anonymous linkage system including secure updates to the salt or seed used in establishing the hashed identifiers. Through a trusted third party the ability to identify who is in what database is technically feasible under appropriate data governance. Large amounts of PII are processed daily through health information exchanges (HIE) and to support Accountable Care Organization operations as well as to processes submitted insurance claims for reimbursement.

B. METHODS FOR DATA LINKAGE

Direct identifiers, as exchanged for treatment, payment, and operations purposes, could theoretically be used to provide direct linkage between CDRN data partners and Sentinel administrative health plan data partners. For example, HIEs are a core goal of the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act[6] that utilize an Enterprise Master Patient Index service to match patients based on the likely match of common patient identifiers (e.g. last name, first name, and date of birth) with the potential to achieve highly accurate matches using algorithms that minimize the need for human review.[7-9] However, some workgroup members indicated that establishing the use of data linkage activities that rely on transfer of unencrypted (or "clear text") patient identifiers for the purposes of research, would be challenging, since sending direct patient identifiers raises privacy and security concerns. However, for some use cases such as where the patient has provided informed consent to perform linkage the use of direct PII identifiers may be permissible.

Privacy preserving methods, such as anonymous linkage strategies through the creation of an anonymous hash identifier through a common hashing algorithm with secure transmission of the salt or seed, could also be employed. A number of PCORnet CDRNs have successfully implemented an anonymous linkage tool using encrypted hashed identifiers to identify overlapping patients across a large urban region[10] and aggregate EHR data for research purposes. This method, which distinctly separates the parties who manage the secret key from the party that receives the hashed IDs is compliant with the HIPAA requirements for a de-identification method and has been piloted within PCORnet at two CDRNs and with one cross CDRN-PPRN linkage. Similar approaches have been adopted by private entities to link administrative claims to EHR data for research.[13]

C. OTHER SOURCES FOR INTEGRATED CLAIMS AND EHR DATA RESOURCES FOR RESEARCH

For the purposes of exploring PCORnet and Sentinel collaborations, this white paper focuses on the technical and governance challenges in the scenario in which data are exchanged across organizational boundaries of disparate data custodians (PCORnet CDRNS, PPRNs, and Sentinel administrative health plan data partners) in order to create a centrally located linked dataset. However, this discussion would not be complete without reference to two additional mechanisms through which integrated data could be available for research or medical product surveillance.

One mechanism to obtain patient-level health data (on a patient-by-patient basis) would be for the patients to access online data profiles through BlueButton Plus[11] or similar patient portals to extract electronic data from either CDRNs or Sentinel administrative health plan data partners following enrollment in a study. The patient would consent to participate in either the trial or observational research and would authorize access to existing patient portals. This would be a less efficient mechanism because it requires every individual patient to extract their data from multiple sources (or authorize access to it possibly through an honest broker) and provide it for the various research and public health activities. The organization(s) receiving these data from patients then link the patient-provided data with their internal data to create a merged dataset. The diversity of patient portal platforms across multiple provider healthcare delivery systems and considerable data and analytic resources required on the part of PCORnet or Sentinel data partners makes this approach impractical. A second mechanism would leverage the large amounts of clinical and claims data that are already being processed daily through health information exchanges (HIEs) and to support Accountable Care Organization operations, or the processes through which insurance claims are submitted for reimbursement. With the appropriate business associate agreements, and requisite patient privacy and data security processes in place, these data could be integrated and used for research or medical product surveillance purposes. Creation of such integrated data assets that can be used for multiple purposes such as research, quality improvement, or medical product surveillance has been actively pursued by private entities, and similar assets could also be leveraged to support PCORnet and Sentinel activities.

VI. ORGANIZATIONAL-LEVEL OVERLAP

Identifying the extent of overlap in patient populations between organizations is a necessary first step to assess the utility of linkage between specific Sentinel administrative health plan data partners and PCORnet data partners for a specific public health or research activity. Since insurers have information on the providers and facilities they have paid, and providers have information on which insurers have paid them for care delivered, preliminary estimates of population overlap for a potential data linkage collaboration could be performed using data available within each data partner, obviating the need to share personally identifying information (PII) to estimate population overlap. Alternatively, privacy preserving methods such as use of hash tables can be applied to estimate organizational overlap.

A. CDRN ASSESSMENT: OVERLAP BASED ON CDRN IDENTIFIED INSURANCE STATUS USE CASE

CDRNs can determine overlap with Sentinel administrative health plan data partners based on their patients' insurance status. They are also able to identify dually insured individuals or track changes in

insurance status, which is not visible to the Sentinel administrative health plan data partners. In the example noted in **Figure 2**, the CDRN entity has the ability to determine insurance status at the time of a clinical encounter and identify the relevant Sentinel administrative health plan data partner(s). Assessing the overlap based on CDRN identified insurance status would help assess the utility of linkage between specific PCORnet data partners and Sentinel administrative health plan data partners.

Although conceptually straightforward, assessing this overlap can be technically demanding since patient insurance information is operationally difficult for some CDRN investigators to obtain. Since patients change insurers with some frequency, the overlap needs to be assessed at the time the information is needed. Because of these challenges, it may be preferable to determine the overlap for specific cohorts that are the subject of a specific inquiry over a defined period, rather than for the entire PCORnet or Sentinel System population.

B. SENTINEL ADMINISTRATIVE HEALTH PLAN DATA PARTNER ASSESSMENT: OVERLAP ASSESSED VIA INSTITUTIONAL PROVIDER IDENTIFIERS USE CASE

A Sentinel administrative health plan data partner can characterize overlap with a CDRN data partner datamart and characterize encounters at hospitals and clinics not affiliated with a CDRN data partner. This can be accomplished by using facility national provider identifiers (NPI), Tax Identification Numbers (TINs) and/or other facility code identifiers submitted with claims to identify members with clinical encounters at hospitals and clinics affiliated with a CDRN data partner. For these patients, it is possible to characterize care obtained outside of individual CDRN provider data partner datamart(s) and highlight care that is not captured within an institutional CDRN datamart. Sentinel administrative health plan data partners can also use this information to identify individuals who receive care from more than one PCORnet CDRN institution or across a CDRN's individual facility datamarts. While the use of provider and institutional identifiers is conceptually straightforward, Sentinel has not yet used these data fields; curation will be needed and there will likely be a learning curve as the data are used.

C. ASSESSMENT OF ORGANIZATIONAL OVERLAP THROUGH USE OF HASHED IDENTIFIERS

Privacy preserving methods such as use of hash tables can also be used for rapid lookup for matches between data sets. One could imagine this as an intermediate step in which potential partners could check whether a patient from a PCORnet CDRN has any claims data within a participating Sentinel administrative data partners' dataset. With this approach, potential linkage data partners would be distributed a universal salt from a trusted third party and create a table of hashed identifiers (i.e., no health information included). These hash tables can then be shared either directly between two potential data linkage partners or with a trusted third party to quantify organizational overlap among multiple potential linkage partners. Once linkage partners are identified, this approach can expanded to support querying and linkage of clinical or claims data for specific studies. A universal salt is not foolproof; it carries some risk of a brute force attack by a determined party in possession of the salt. This risk could be mitigated by the clear separation of access to the encryption key (salt and hashing) algorithm and the resultant hashes, by encryption of the salt itself, and by changing the salt each time data are refreshed.

VII. DATA LINKAGE USE CASES

While characterization of the potential overlaps in patient populations at the organizational level can identify promising partnerships for building cohorts to support research or medical product safety surveillance, the linkage of information from disparate sources requires reaching agreement on governance as well as addressing a number of technical aspects to ensure data security, patient privacy, compliance, and other protections. The workgroup identified key differentiating attributes that seek to inform data collaboration needs in terms of their distinct challenges and considerations with regard to data linkage. The **Appendix** outlines an example process for technical transfer of data and serves as a guide to the governance discussion presented below.

A. TYPOLOGY OF PATIENT-LEVEL LINKAGE COLLABORATIONS

Key attributes of patient-level linkage collaborations include the type of activity (medical product safety surveillance or research), patient privacy and informed consent, and data linkage frequency. **Table 1** provides a typology of patient-level linkage studies, characterized according to key attributes that have distinct considerations in terms of data collaboration needs. First, determination of an activity as research or as a medical product safety surveillance activity will be associated with distinct requirements. Sentinel administrative health plan data partners, including participating PCORnet CDRNs, must adhere to the privacy provisions of the Health Insurance Portability and Accountability Act (HIPAA) as described for public health activities for all Sentinel activities. In addition, the Office for Human Research Protections has determined that the regulations administered by that office (45 CFR part 46) do not apply to activities that are included in FDA's Sentinel Initiative. Therefore, the work of Sentinel is not reviewed by Institutional Review Boards (IRB). For observational and randomized CER originating out of PCORnet or other funders, these activities would be considered human subjects research and would require appropriate IRB oversight. Activities that are categorized as research would require compliance with both HIPAA and the Common Rule (45 CFR part 46). To link data from multiple data sources for research purposes, patient informed consent, or waiver of such consent and a waiver of HIPAA authorization will be required.

Access to PII such as Social Security Numbers (SSNs) varies across CDRNs and PPRNs, as well as across sites within a single CDRN. In addition, the ability and willingness of PCORnet and Sentinel partners to share direct patient identifiers, e.g., CDRN-to-health plan or vice versa, depends in part on whether a patient has provided informed consent that authorizes linkage of their clinical data to healthcare claims or a waiver of informed consent and HIPAA authorization is in place through an established data use agreement or business associate agreement. Some workgroup members indicated that approval and ability to share PII between partners or with the Coordinating Center or a trusted third party would be more likely in cases where patient informed consent has been obtained and mechanisms are in place to collect this information directly from the patient. There are significant barriers in establishing linkage for observational research within many organizations. For example Illinois law requires that patients have opt-out privileges for PHI or PII used for research outside of the provider system that collected the data. This mandates the use of anonymous linkage strategies to fully comply with the creation of a deidentified dataset. Finally, refreshes and updated data will be required over time to address data collaboration needs of research studies and surveillance activities (versus activities requiring one-time data linkage).

B. USE CASES

For each type of activity, the workgroup relied on use cases to guide a more detailed discussion and develop recommendations. Use Case 1 through 3 describe scenarios in which data linkages are performed on a project-by-project basis. Use Case 4 describes a scenario in which a persistent linked dataset is created to support multiple purposes.

For Use Case 1 through 3 (project specific linkages), the privacy-preserving approaches described for assessing organizational overlap can be extended to act as an underlying infrastructure in which querying and linkage of data can be performed efficiently for multiple projects. Several CDRNs have adopted an approach to anonymous linkage software[10] which limits central data aggregation and instead keeps data at the source locations and stores a central set of hash tables. This model follows similar guidelines for encryption and linking identifiers from multiple organizations, but linkage of the individuals' data is performed on demand. Specifically, in this method all clinical data resides locally at each institution until there is a specific use that requires creation of a linked dataset. A trusted third party creates non-derived linkage identifiers on a project-by-project basis after receiving the salted hashes from the participants but does not itself store any clinical data. For additional security, the encryption keys themselves can be encrypted and kept separate from the hub which receives the initial linkage identifiers. Participants may then share data with the non-derived linkage identifiers with a fourth party (e.g., the project coordinating center) that will conduct the data linkage; in this case, since the encryption key identified by the third party is kept separate from the data partners and the fourth party conducting the data linkage, the patient is further protected. This "on-demand" linkage approach was supported by several workgroup members as an intermediate step between development of linkage collaborations and processes on a project-by-project basis and creation of a persistent linked dataset.

1. Use Case 1: Patient Provides Consent

For **Use Case 1: Patient Provides Consent**, the workgroup developed recommendations and considerations related to patient-level linkage studies when the patient provides consent that includes authorization to payers to provide data to investigators. The workgroup used the Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-term Effectiveness (ADAPTABLE) trial as the worked example in developing the recommendations.

a. Institutional Review Board approval

Most workgroup members agreed that the use of a central IRB, deferral or reliance on another IRB would be preferable. At the time of this writing, a Notice of Proposed Rule Making to revise the Common Rule calls for the use of a single IRBs; this approach may thus become a standard. The signed consent would give investigators authorization to conduct the data linkage.

b. Patient population available for data linkage

For some Sentinel administrative health plan data partners, patient informed consent may not be sufficient to allow provision of at least certain subsets of patient data by Sentinel DP to CDRN. For example, health insurer partners may be restricted in their ability to access PII for this purpose. An example is the situation in which the health insurer is providing administrative services to a self-insured employer. Although individuals own their own medical records, insurance claims may be considered the

property of the purchaser. Some employers may be more or less amenable to such sharing of data on a case by case basis. Seeking modifications to business agreements to allow use of data for research would need to be pursued on a case-by-case basis and while possible, would be highly resource intensive. Efforts are underway at several Sentinel administrative health plan data partners to revise the business agreements and provide for research provisions.

c. Content of data provided

Workgroup members agreed that the data shared should be limited to specific individuals, period, and data elements necessary for the conduct of each study, to conform to a "minimum necessary" principle. For example, the ADAPTABLE trial will compare the benefits and harms of a low- and regular-strength daily dose of aspirin in patients diagnosed with heart disease, the trial will require limited data on enrolled subjects to ascertain cases from Sentinel administrative health plan data partners.[12] In this use case the ADAPTABLE trial is interested in the longitudinal follow-up identified through administrative claims and the capture of events that occur outside of participation in the CDRN. For Medicare beneficiaries, ADAPTABLE investigators intend to obtain individuals' claims data using extant policies of the CMS Research Data Assistance Center (Res DAC).

d. Data linkage and data flow requirements

With informed consent that specifies with whom the PII would be shared and stored, DUAs, and a secure data transfer portal in place, CDRNs indicate that the transmission of the PHI data to a Coordinating Center to conduct the trial is straightforward since the patient can be asked for the necessary PII identifiers. The patient would thus provide both the identifiers and the permission to use the identification in the proposed linkage activities. The technical steps to utilize direct patient PII from CDRNs to respective health insurers for linkage may not be straightforward. The complexity of the United States health insurance marketplace often has many named plans represented under a single health insurance health plan, which may result in incomplete documentation of the patient's insurer in EHRs; integration with CDRN financial systems will be critical in obtaining validated insurance plan and health plan identifiers. Participating subjects may change health plans during the trial. Additionally, when subjects withdraw informed consent health plan data partners would need to be notified to discontinue the transfer of data on a particular patient in subsequent study data refreshes.

2. Use Case 2: Waiver of Consent and HIPAA Authorization

For **Use Case 2: Waiver of Consent and HIPAA Authorization**, the workgroup developed recommendations and considerations related to patient-level linkage studies when the patient is not consented and a waiver of informed consent is sought from an IRB and a waiver of HIPAA authorization is sought from a privacy board. Such approval has been provided in the past, principally on the basis of need for the information (restriction to patients who provide consent would introduce unknown biases and impair generalizability of the results), limited risk to patients, and impracticality of obtaining consent. The workgroup used the PCORnet bariatric surgery cohort and the pediatric antibiotic weight studies as the worked example in developing the recommendations.

a. Institutional Review Board approval

The issues differ from use case 1 as the investigators do not have patient consent and authorization, thus requiring IRB approval of a waiver of informed consent and a HIPAA waiver of authorization to implement the study. Most workgroup members agreed that, if feasible, use of a central IRB or ability to defer to another IRB would be preferable to seek a waiver of consent, to improve study efficiencies. As in use case 1, all participating partners, and for CDRN data partners without a central IRB all participating entities within a CDRN, should be listed as sites in the IRB application. At least one partner indicated that it would likely not be able to rely on a central IRB, given the specific review requirements associated with use of their data for research. As of this writing, there is a notice of proposed rulemaking to revise the Common Rule; this revision calls for increased use of central IRBs, and so this particular limitation may not persist. As it is not feasible to obtain individual informed consent and authorization in a large observational longitudinal study, a waiver of consent and HIPAA authorization would need to be sought. Several patient and community groups raised concern regarding such waivers.

b. Patient population available for data linkage

For some health insurer data partners, IRB approval and a waiver of informed consent may not be sufficient to allow provision of at least certain subsets of patient data by Sentinel DP to CDRN. For example, health insurer partners may be restricted in their ability to access PII to identify individuals if an existing business agreement with an employer or provider prohibits use of their data for research, even with authorization from the patient. Seeking modifications to business agreements to allow use of data for research would need to be pursued on a case-by-case basis and, while possible, would be highly resource intensive.

c. Content of data provided

The bariatric surgery studies will need to obtain adverse outcome events related to bariatric surgeries including re-hospitalization and death that are likely to happen outside the PCORnet data partners and would require collaboration with Sentinel data partners and mortality data files (either the Social Security Death Index or the National Death Index). The antibiotic study has need to obtain antibiotic dispensing data from Sentinel administrative health plan data partners, as data on both prescriptions written by PCORnet institutions as well as antibiotic prescriptions written at clinical sites outside of PCORnet and subsequently dispensed are required.

d. Data linkage and data flow requirements

The workgroup agreed that it will be more challenging to share PII without informed consent that specifies with whom the PII would be shared and stored. With a waiver of informed consent and HIPAA authorization, DUAs, and a secure data transfer portal in place, some participating CDRNs indicated that the transmission of PII data to a Coordinating Center to conduct the study would be possible. There are CDRN organizations that have indicated that no PHI or PII can be removed from their source systems due to more restrictive laws within their states. As such the anonymous linkage hashing algorithm approach would be the only feasible mechanism to securely provide data linkage activities. The technical steps to appropriately identify direct patient identifiers and health insurance information would require that PCORnet data partners obtain accurate information from the financial systems within local CDRN

datamart(s). Further complicating identification of unique insurance identifiers will be patients who have multiple sources of insurance or who may switch insurance during the studies. This information would change overtime and would require the governance and oversight activities to keep these sources up to date at each refresh as unique identifiers would change over defined periods.

3. Use Case 3: Public Health Activity

For **Use Case 3: Public Health Activity**, as determined by FDA under its public health authority, the workgroup developed recommendations and considerations related to a patient-level linkage analysis for medical product safety surveillance in which the population is identified by Sentinel administrative health plan data partners as part of an FDA public health activity. The PCORnet CDRNs would be covered under the FDA Sentinel medical product surveillance as data partners. The use case would require additional clinical INR values from CDRN sites in a protocol-based assessment of dabigatran:

a. Institutional Review Board approval

As this use case arises as part of a FDA Sentinel public health activity, IRB approval would not be necessary to conduct the evaluation.

b. Patient population available for data linkage

Sentinel data partners would identify the new users of warfarin and dabigatran. Among the study population, Sentinel partners would then identify patients with medical encounters of interest at a PCORnet institution based on the NPI, TIN, and/or other facility code identifiers submitted with the claims or via hashed identifiers if these are available, and additional clinical information for these patients would be requested from the PCORnet partners.

c. Content of data provided

Additional INR values available at CDRN sites for patients identified would need to be transferred to Sentinel administrative health plan data partners for inclusion in routine data analytic code developed for medical product safety surveillance.

d. Data linkage and data flow requirements

Sentinel data partners would need a secure mechanism to exchange individual-level data. A first decision would involve whether the exchange uses direct identifiers or anonymous linkage identifiers. When multiple CDRNs need to link to multiple Sentinel data partners, the Coordinating Center (Sentinel or PCORnet) or another third party could pass identifiers from the requesting organizations to the receiving organizations and vice versa, potentially through anonymous linkage, to the CDRNs to obtain the minimal data necessary to address the public health evaluation. Using the Coordinating Center or another intermediary could have the advantage of avoiding many separate pairwise exchanges between requesting and receiving organizations.

4. Use Case 4: De-identified Linked Dataset for Multiple Purposes

For **Use Case 4, De-identified Linked Dataset for Multiple Purposes**, the workgroup identified considerations for the creation of a standing de-identified linked dataset that could be used to address multiple questions, some of which may not be identified at the time of the data linkage. The existing OptumLabs collaboration[13] is an example of such a standing linked dataset sourced from multiple participant organizations. Several key elements of this model include: use of an encryption technology that minimizes the risk of re-identification; centralized hosting of the statistically de-identified linked datasets in a secure data warehouse at a trusted party; development of extensive privacy and data security measures; and ongoing engagement of the various participants to guide the development of a research agenda to ensure that research priorities and allowable uses of the linked resource are aligned with participant organizational needs.

A number of workgroup members indicated that creation of a standing linked data resource would create the most efficient resource to support rapid feasibility queries, response to medical product safety surveillance, and comparative effectiveness research. Since PCORnet and Sentinel sites do not date shift and some are geographically distinct, it would be necessary to quantify the risk for re-identification for central standing datasets derived from these organizations.

The acceptability of this approach to PCORnet and Sentinel organizations was not explored. The workgroup acknowledged that governance questions common to all the use cases must first be addressed (e.g., where would the linked dataset be stored? what variables would be included in the linked dataset? what patient privacy and data security processes will be put in place to minimize the risk for breach or inadvertent disclosure?). The use case of a standing linked dataset also raises additional considerations. These include the need for clearly defined allowable uses for the linked dataset, resolution on ownership and access, development of technical processes and procedures to support updating of the data and data linkages, and the creation of a committee with broad representation to provide ongoing oversight on governance. Clarity on a shared sense of mission and establishment of mutually agreed upon governance that addresses organizations' varied concerns and needs will be critical to successfully engaging partners in contributing data to a standing linked data resource.

Finally, the workgroup discussed collaborations with PPRNs and surveyed the PPRN leads about the availability of insurance information for the potential of conducting linkage to their membership. There was limited availability of insurance information within most PPRNs. As work evolves and specific use cases are presented, further discussions on data linkage with Sentinel administrative health plan data partners will be required. Similar strategies to the use cases presented above could be employed with PPRNs.

VIII. DISCUSSION

Building successful collaborations between Sentinel administrative health plan data partners and PCORnet data partners to link these complementary data sources requires agreement on both governance and technical issues to ensure patient privacy, data security, compliance, and proprietary needs are met. While both will require effort, the technical work can build on the fact that clinical facilities and health plans routinely perform closely related activities in the course of normal business operations. Many of the governance issues will require a combination of policy that establishes parity

between the data partners whose data are being linked to assure robust security for the linked data and establish clear rules for its use. From a sustainability perspective the creation of multiuse linked data resources is desirable. However, their creation, maintenance, and rules of use present additional governance challenges to address. Many of these issues will be more tractable if the initial focus is on answering specific questions, thus limiting both the population and the specific data elements that are shared. Addressing the data linkage issues in this stepwise fashion will allow incremental development of policies, procedures, and infrastructure. It will also foster the development of trust and a shared sense of mission among the collaborating data partners.

We therefore advocate a stepped approach that allows identification of the size of a shared population, without disclosure of any patient-level data. A substantial piece of this activity can actually be performed by either the PCORnet CDRNs or the Sentinel administrative health plan data partners without the participation of the other. Collaboration in this activity can provide clear guidance about the PCORnet CDRN data partners' contribution to their datamarts, and the Sentinel administrative health plan data partners can characterize the distribution of care that CDRN patients receive in organizations that do not contribute to the PCORnet CDRNs' datamarts. This organizational-level of overlap will highlight the missed clinical encounters between the organizations, as PCORnet lacks information about care obtained outside of the CDRN datamart(s) and Sentinel large administrative health plan data partners lack the depth of clinical care obtained at PCORnet CDRNs. We note that even this level of data sharing raises concerns about disclosure of potentially sensitive information regarding PCORnet CDRN institutions' practice patterns and Sentinel administrative health plan data partners' local markets.

The simplest use case for actual data sharing involves having the Sentinel administrative health plan data partner provide information about specific health events in response to members' direct requests, as in the ADAPTABLE use case. Follow-up of clinical trial outcomes is an example of this kind of sharing. In this example the combination of several features minimizes the challenges. These include the members' authorization, the limited amount of information needed, e.g., dates of hospitalization for acute myocardial infarction, and the fact that the data is transferred to a Coordinating Center with robust data security. Health plan members would request sharing through the provision of a signed informed consent document that would authorize the Sentinel administrative health plan data partner to release the requested PHI to the CDRN site for analysis.

The next most challenging scenario involves similarly limited data sharing, but in the setting of an observational study for which there is no individual authorization to either party to share data. We describe an example of work of this type conducted under HIPAA's provisions for public health activities (as applies to the Sentinel System's activities on behalf of FDA) and other examples of research conducted with an IRB's waiver of informed consent as allowed by the Common Rule. The use of either direct identifiers or an anonymous hashed identifier can work in this setting. However, at some CDRN sites anonymous linkage appears to be the only approach permissible to engage in the creation of a linked dataset for medical product safety surveillance or observational research activities. The use of an anonymous hashed identifier to create a persistent database of what data resource individuals contribute to would facilitate identification for linkage activities. This would be an intermediate step in the creation of a fully linked dataset to support multiple activities.

Potentially the most valuable, but also the most challenging scenarios will be ones in which a large linked dataset is created to support an array of evaluations. As discussed the use of anonymous linkage

through hashed identification is appealing. However, significant infrastructure would need to be put in place to utilize a common method across multiple organizations coordinated with refreshed data and the need to keep the salt or seed character secure through a trusted party. There can be great value in a query-ready resource with linked, quality checked data that can rapidly address a wide array of medical product safety surveillance and comparative effectiveness research questions as they arise. Creating such a resource is technically feasible, although it is considerably more challenging than creating a single use dataset, because of the need to refresh the data on a regular basis. However, precisely because of the many potential uses of such a resource, and because of the risks of unauthorized use, or unintended disclosure in a large linked resource, it is also considerably more challenging to develop the governance policies for such a resource.

Although beyond the scope of this document, we note that there is important ongoing work in developing methods for analyzing vertically partitioned data without sharing of individual level data.[14,15] In these cases, only an individual identifier, which can be encrypted, is shared. While these methods are not currently usable for the kinds of examples described in this document, they may become available within the next few years. These privacy protecting methods would mitigate many, but not all, of the challenges described here.

In summary, there are compelling reasons to develop linkage policies, procedures, and infrastructure to allow use of the complementary data sources in claims and EHR data. We propose an incremental approach to addressing the technical and governance challenges in data linkage infrastructure.

IX. REFERENCES

1. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood). 2014;33(7):1178-1186.
2. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:23-31.
3. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21(4):578-582.
4. 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(4):583-586.
5. Devoe JE, Gold R, McIntire P, Puro J, Chauvie S, Gallia CA. Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. Ann Fam Med. 2011;9(4):351-358.
6. Blumenthal D. Launching HITECH. N Engl J Med. 2010;362(5):382-385.
7. Clayton PD, Narus SP, Huff SM, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med. 2003;42(1):1-7.
8. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. A practical method of linking data from Medicare claims and a comprehensive electronic medical records system. Int J Med Inform. 2003;71(1):57-69.
9. Arellano MG, Weber GI. Issues in identification and linkage of patient records across an integrated delivery system. J Healthc Inf Manag. 1998;12(3):43-52.
10. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc. 2015;22(5):1072-1080.
11. D'Amore JD, Mandel JC, Kreda DA, et al. Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative. J Am Med Inform Assoc. 2014;21(6):1060-1068.
12. Hernandez AF, Fleurence RL, Rothman RL. The ADAPTABLE Trial and PCORnet: Shining Light on a New Research Paradigm. Ann Intern Med. 2015.
13. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: building a novel node in the learning health care system. Health Aff (Millwood). 2014;33(7):1187-1194.
14. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. J Am Med Inform Assoc. 2013;20(3):453-461.
15. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc. 2012;19(5):758-764.

X. APPENDIX: EXAMPLE PROCESS FOR TECHNICAL TRANSFER OF DATA

The example presented below is for a complex data linkage scenario in which more than one data partner (CDRNs or Sentinel administrative health plan data partners) is requesting data from multiple data partners. The following nomenclature is used to describe the data transfer process: (1) "requesting data site" represents the institution requesting patient-level CDM data, (2) "coordinating center" represents a data hub through which data from requesting and receiving data sites is transmitted securely, (3) "receiving data site" represents the institution receiving the request to supply patient-level CDM data, (4) "arbitrary patient identifier" is a data partner specific identifier that is not a medical record number or unique member number, and (5) "study ID" is a Coordinating Center derived identifier to limit the transmission of identifiable data. Through the use of local arbitrary identifiers and Study_ID, the transmission of patient-level identifiers is reduced to two steps, from requesting data site to the Coordinating Center and from Coordinating Center to the receiving data site. This process is provided for illustrative purposes only; creation of a standard set of operating procedures would require obtaining legal, compliance, and other organizational approvals.

1. Requesting data site(s) sends their patient listing to the Coordinating Center, which includes the variables to be used for linkage along with an arbitrary ID (stores crosswalk in local data partner study folder)
    - Patient identified data could be supplied as an anonymous hashed identifiers with all data partner participants utilizing the same software and hashing criteria
    - Coordinating Center will need to be able to translate insurance names into respective Sentinel data partner names
    - Coordinating Center will need to be able to translate provider ID (NPI) to respective CDRN data partner datamarts.
2. Coordinating Center assembles the patient lists from requesting data sites and assigns a Study_ID to each patient
3. Coordinating Center subsets the patient lists into lists for specific receiving data site(s) and sends the patient listing to respective receiving data site(s)
    - Patient identified data could be supplied as an anonymous hash with all data partner participants utilizing the same software and hashing criteria
4. Receiving data site(s) uses the patient listing to identify patients in their internal data sources and develops a crosswalk mapping the arbitrary Pat_ID to the patient identifier used in the Sentinel or PCORnet CDM. This crosswalk is stored and remains in the local data partner study folder.
5. Receiving data site(s) sends the minimum CDM data necessary to address the data request using the Study_ID
6. Coordinating Center assembles data from receiving data sites
7. Coordinating Center distributes a data file to the respective requesting data site(s) data using arbitrary_ID supplied in step 1

302-547-8167 8101

A. EXAMPLES

1. ADAPTABLE Data Request from CDRNs

Requesting Data Sites: CDRNs
Receiving Data Sites: Sentinel Data Partners

Step 1: Patient listings are sent from requesting data sites (CDRNs) to the Coordinating Center
From requesting data site: CDRN1
Arbitrary_ID, Name, DOB, Insurance, Insurance ID, request_ds_name
11, Mickey Mouse, MM/DD/YYYY, Health Plan 1, Insurance ID, CDRN1
…
From requesting data site: CDRN 2
Arbitrary_ID, Name, DOB, Insurance, Insurance ID, request_ds_name
33, Donald Duck, MM/DD/YYYY, Health Plan 2, Insurance ID, CDRN2
…

Step 2: Patient lists are consolidated at the Coordinating Center
Study_ID, Arbitrary_ID, Name, DOB, Insurance, Insurance ID, request_ds _name
1, 11, Mickey Mouse, MM/DD/YYYY, Health Plan 1, Insurance ID, CDRN1
2, 33, Donald Duck, MM/DD/YYYY, Health Plan 2, Insurance ID, CDRN2

Step 3: Coordinating Center generates patient listings for specific receiving data sites (Sentinel administrative health plan data partners) and sends the patient lists to the respective receiving data site(s)
To receiving data site: Sentinel DP1
Study_ID, Name, DOB, Insurance, Insurance ID
1, Mickey Mouse, MM/DD/YYYY, Health Plan 1, Insurance ID
…
To receiving data site: Sentinel DP2
Study_ID, Name, DOB, Insurance, Insurance ID
2, Donald Duck, MM/DD/YYYY, Health Plan 2, Insurance ID
…

Step 4: Receiving data site(s) match patients
Use SAS list of patients supplied by Coordinating Center to extract minimum necessary CDM into DPLocal record of supplied member list. This step requires mapping with raw untransformed data at data partner site(s) to map to arbitrary CDM identifier.

Step 5: Receiving data site(s) extract minimum data necessary
Use SAS programs supplied by the Coordinating Center in request to narrow returned data to minimum necessary to address research question; populate Study_ID with Study_ID supplied from Coordinating Center; leave crosswalk file in DPLocal

Step 6: Coordinating Center assembles returned data from receiving data site(s)

Step 7: Coordinating Center returns individual data to requesting data site(s)

2. Dabigatran Data Request from Sentinel

Requesting Data Sites: Sentinel Data Partners
Receiving Data Sites: CDRNs

Step 1: Patient listings are sent from requesting data sites (Sentinel administrative health plan data partners) to the Coordinating Center
From requesting data site: Sentinel DP1
Arbitrary_ID, Name, DOB, receiving_ds_name, request_ds _name
77, Mickey Mouse, MM/DD/YYYY, CDRN1, Sentinel DP1
…
From requesting data site: Sentinel DP2
Arbitrary_ID, Name, DOB, receiving_ds_name, request_ds _name
88, Donald Duck, MM/DD/YYYY, CDRN2, Sentinel DP2
…

Step 2: Patient lists are consolidated at the Coordinating Center
Study_ID, Arbitrary_ID, Name, DOB, receiving_ds_name, request_ds _name
1, 77, Mickey Mouse, MM/DD/YYYY, CDRN1, Sentinel DP2
2, 88, Donald Duck, MM/DD/YYYY, CDRN2, Sentinel DP2

Step 3: Coordinating Center generates patient listings for specific receiving data sites (CDRNs) and sends the patient lists to the respective data site(s)
To CDRN1:
Study_ID, Name, DOB
1, Mickey Mouse, MM/DD/YYYY
…
To CDRN2:
Study_ID, Name, DOB
2, Donald Duck, MM/DD/YYYY
…

Step 4: Receiving data site(s) match patients
Use SAS list of patients supplied by Coordinating Center to extract minimum necessary CDM into DPLocal record of supplied member list. This step requires mapping with raw untransformed data at data partner site(s) to map to arbitrary CDM identifier.

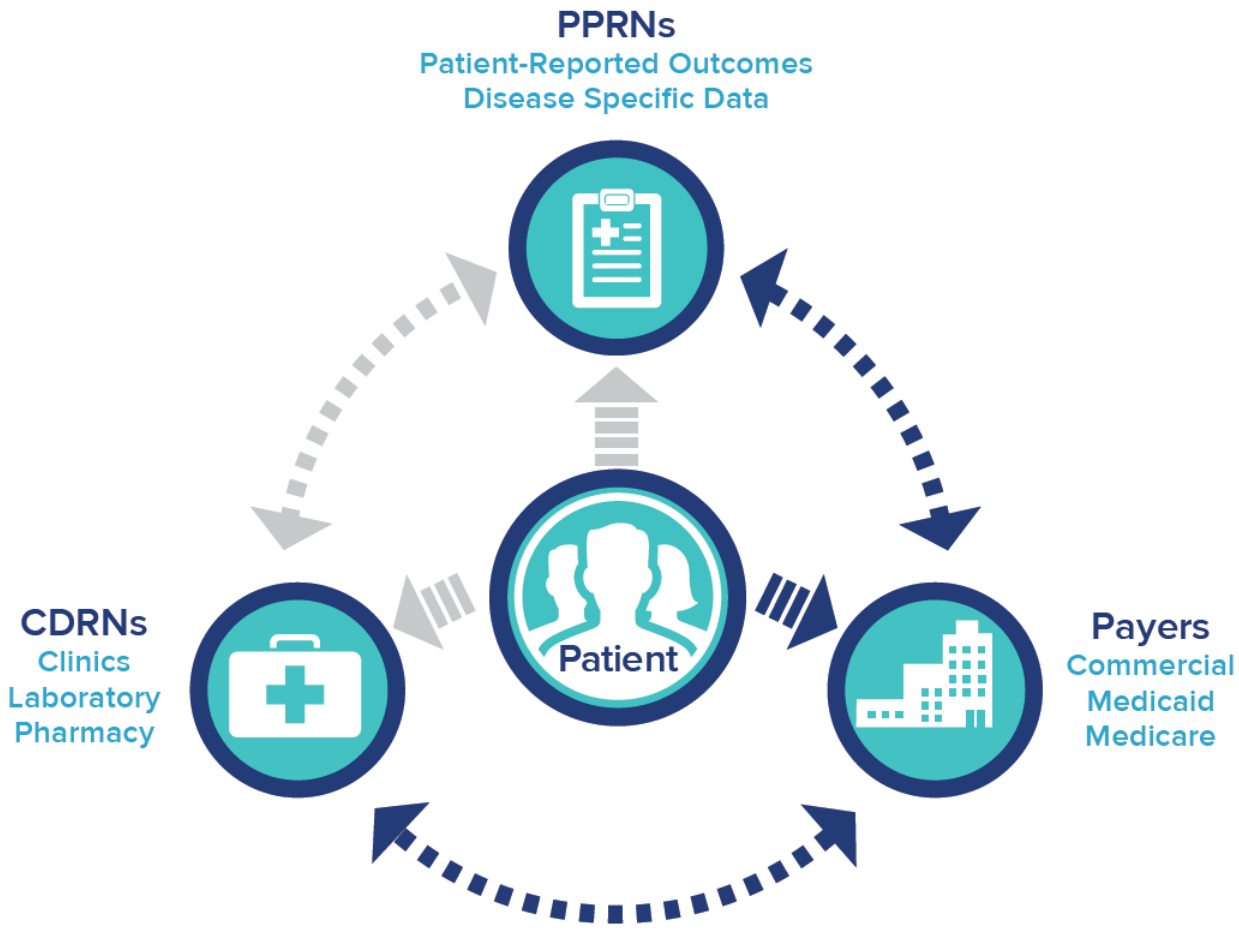
Step 5: Receiving data site(s) extract minimum data necessary
Use SAS programs supplied by the Coordinating Center in request to narrow returned data to minimum necessary to address research question; populate Study_ID with Study_ID supplied from Coordinating Center; leave crosswalk file in DPLocal

Step 6: Coordinating Center assembles returned data from receiving data site(s)

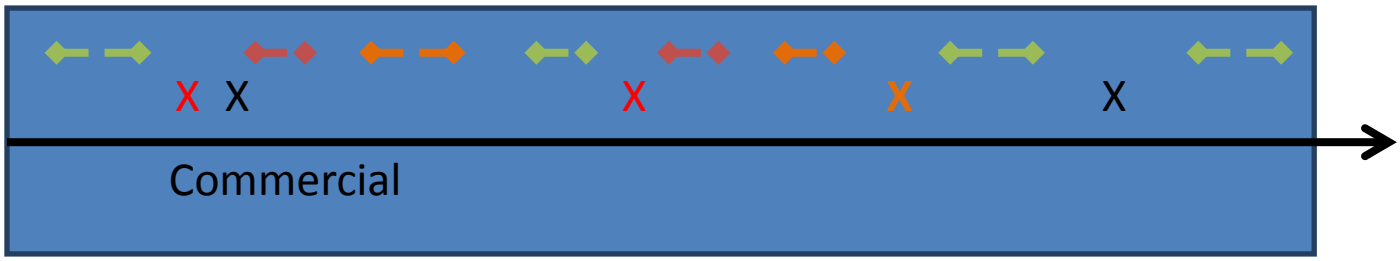
Step 7: Coordinating Center returns individual data to requesting data site(s)

XI. TABLES & FIGURES

A. FIGURE 1. CONCEPTUAL MODEL



B. FIGURE 2. WORKING EXAMPLE



Local hospitalization

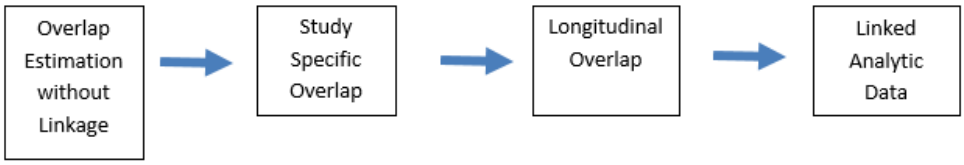
CDRN entity A hospitalization

CDRN entity B hospitalization

X Other clinical outpatient encounters (e.g.

X Pharmacy

C. FIGURE 3. SPECTRUM OF COLLABORATION



D. TABLE 1. TYPOLOGY OF PATIENT-LEVEL LINKAGE STUDIES

Study design	Research or public health activity?	IRB approval needed?	Patient informed consent?	PCORnet - Sentinel DP data refresh and re-linkage needed?	Description of use case
Randomized controlled trial (RCT), prospective follow-up	Research	Yes	Yes	Yes	ADAPTABLE RCT
Observational cohort study, one-time data linkage*	Research	Yes	Waiver of informed consent sought from IRB and HIPAA from Privacy Board	No	Bariatric surgery outcomes study; Pediatric antibiotics study
Observational cohort study, prospective follow-up*	Research	Yes	Waiver of informed consent sought from IRB and HIPAA from Privacy Board	Yes	(Deferred)
Creation of a single cohort with potential multiple uses*	Research	Potentially yes	Potentially yes	Yes	OptumLabs collaboration
Observational cohort, evaluation, one-time data linkage*	Medical product safety surveillance activity	No	Not applicable	No	FDA Mini-Sentinel custom protocol study example
Observational cohort, surveillance with prospective follow-up*	Medical product safety surveillance activity	No	Not applicable	Yes	FDA Mini-Sentinel active surveillance example

*If new research aims are identified, a protocol modification and IRB amendment would be sought. CDRN, PPRN, and Sentinel administrative health plan data partners, as appropriate, would have the opportunity to "opt in" to participate in the amended scope of inquiry, pending receipt of appropriate internal review approvals.

E. USE CASE 1

Use Case 1: Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-term Effectiveness (ADAPTABLE Aspirin Study)

Research Activity: Patient provided informed consent with prospective follow-up

Primary Aim: To compare the effectiveness of two daily doses of aspirin in reducing death and hospitalization for heart attacks and strokes in a secondary prevention population of patients with atherosclerotic cardiovascular disease (ASCVD).

Study Design: Prospective, patient-level randomized controlled trial of high risk patients who have had a heart attack or have significant blockage of their coronary arteries. Patients will be randomized in a 1:1 fashion to instructions to take an aspirin dose of either 81 mg or 325 mg daily.

Cohort identification and data collection approach: EHR data from CDRNs will be used to facilitate patient recruitment into the trial and to generate study cohorts. Data collection will include use of existing electronic health records from consenting patients in accordance with privacy and security measures, as well as a web-based patient portal to collect patient-reported outcomes (PROs) and data gathered during patients' visits with their clinicians to supplement/support the EHR. Further approaches that may be used to ascertain "out-of-network" outcome events include supplementation with claims data, e.g., via the Centers for Medicare and Medicaid Services (CMS) for CMS-eligible patients or via Sentinel health plan data partners; ideally, refreshes of the claims data would occur on a semi-annual basis for up to two years.

F. USE CASE 2

Use Case 2: Weight Studies

Research Activity: Waiver of patient informed consent and HIPAA authorization with retrospective data collection

Primary Aim: Assess outcomes of different types of bariatric surgery, and, separately of early life exposures to antibiotics

Study Design: Retrospective observational cohort studies

Cohort identification and data collection approach: CDRN sites identify the patients for the study. The CDRN sites would need to be able to identify insurance identifiers from information systems, which are likely contained in financial systems. Direct patient identifiers would need to be shared with a Coordinating Center to be assembled from across the CDRNs to be sent to Sentinel data partners. Given the need for direct identifiers it would be important that the CDRNs can identify which Sentinel data partner the patient belongs to at any given period. Sentinel data partner would supply individual level data back to the Coordinating Center.

G. USE CASE 3

Use Case 3: A protocol for assessment of dabigatran

Public Health Activity: Observational cohort, surveillance with prospective follow-up

Primary Aim: To compare safety outcomes in adults with atrial fibrillation who are new users of dabigatran or warfarin therapy

Study Design: Prospective follow-up of new users of both warfarin and dabigatran

Cohort identification and data collection approach: Sentinel data partners would identify adult patients with atrial fibrillation who are new users of dabigatran or warfarin in their Sentinel CDM. Lists of identified cohort members would need to be supplied either as direct identifiers or through anonymous linkage to PCORnet CDRNs to obtain additional outpatient INR values available at PCORnet CDRNs. This would enable the Sentinel administrative health plan data partners to provide stratification of warfarin users by INR values. Additionally, partnership with PCORnet CDRNs may allow for Sentinel administrative health plan data partners to obtain more detailed clinical data on instructions to patients regarding warfarin therapy.

H. USE CASE 4

Use Case 4: De-identified linked dataset available for multiple purposes

Design: OptumLabs, an open collaborative research and innovation center, was established in 2013 with the primary goal of improving patient care and value in the healthcare system through the development of multi-institution collaboratives.

Patient Population: Determined based on mutual agreement among collaborators

Content of Data to be Linked: Determined based on mutual agreement among collaborators

Data Linkage and Data Flow Requirements: A key asset of OptumLabs is an integrated database which contains de-identified claims and/or clinical data for over 150 million people; the claims data can also be combined with other large research databases contributed by collaborative participants. Participants are provided by the trusted party in OptumLabs with a "salt" code that is universal to the participants with data to be linked. Via a salt and hash process, participants create encrypted (salted and hashed) identifiers to be used for linkage. De-identified datasets that include health information (e.g., administrative claims data, EHR data, and/or patient-generated data) and the salted-and-hashed identifiers are provided by the participants to the trusted party to perform the data linkage. After the data are linked, the trusted party applies a second salt and hash to further de-identify the data elements in order to reduce the likelihood of re-identification. The resulting merged dataset, statistically de-identified in compliance with HIPAA standards, is stored in a secure location separate from the original data used for linkage, and researcher access is provided via secure enclaves (distinct research environments with firewalls) configured to contain only the specific de-identified data required for a given study. As an additional level of protection, researchers are not allowed to download any patient-level data from the enclave. This is intended to prevent unauthorized linkages between the warehoused data and external databases.