PharmaSUG 2015 - Paper HA07

Distributed data networks: A paradigm shift in data sharing and healthcare analytics

Jennifer R. Popovic, Harvard Pilgrim Health Care Institute/Harvard Medical School

Boston, MA

ABSTRACT

Administrative claims data are rich sources of information that are used to inform study topics ranging from public health surveillance to comparative effectiveness research. Data sourced from individual sites can be limited in their scope, coverage and statistical power. Sharing and pooling data from multiple sites and sources, however, present administrative, governance, analytic and patient privacy challenges.

Distributed data networks represent a paradigm shift in healthcare data sharing and are evolving at a critical time when 'big data' and patient privacy can introduce competing priorities. A distributed data network is one for which no central repository of data exists. Rather, data are maintained by and reside behind the firewall of each data-contributing partner in a network, who transform their source data into a common data model and permit indirect access to those data through the use of a standard query approach. Transformation of data to a common data model ensures that standardized applications, tools and methods can be applied to them.

INTRODUCTION

This paper introduces the concept of a distributed data network, its purposes and benefits and, using the Mini-Sentinel pilot project as a case study, discusses using SAS to design and build infrastructure for a successful multi-site, collaborative distributed data network. Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to create an active surveillance system - the Sentinel System - to monitor the safety of FDA-regulated medical products. This paper also describes the SAS-based, open-source analytic system built and maintained by the Mini-Sentinel Operations Center (MSOC).

WHAT IS A DISTRIBUTED DATA NETWORK

A distributed data network is one in which no central repository of data exists. Rather, data are maintained by and reside behind the firewall of each data holder, which allow indirect analytic access to their patient-level data via programming code that is securely distributed to them and intended to execute on their side of the firewall. The data are therefore 'distributed' due to the lack of centrality.

Distributed data networks exist by a set of guiding principles:

- Data holder sites maintain control over their data,
- Data holder sites have standardized their data to a common data model,
- Data holder sites' ongoing involvement is needed in order to interpret data and findings; they know their data the best, so are true ***partners*** in the network (indeed, the terms 'data holder' and 'data partner' will be used interchangeably in this paper),
- Programming code gets securely distributed to data holders for them to execute locally and in a manner that makes it easy for them to execute,
- Following execution of programming code, data holders return results that were produced by the executed code, to the requestor. Typically, data returned are aggregated rather than patient-level.

PURPOSE OF A DISTRIBUTED DATA NETWORK

Distributed data networks often allow for access to more data than what a single or centralized site might be able to offer. By pooling resources (data) across several sites, with security in place such that each site maintains ownership over its own data, these networks provide several key benefits, including:

- Offering alternative ways to study occurrences of rare outcomes, uptake or usage of new drugs or therapies, and diverse populations of individuals,
- Achieving greater statistical power due to larger numbers of observations,
- Encouraging the development of novel analytic and statistical methods that do not rely solely on the use of patient-level data,

- Challenging analytic programmers to approach projects with the intention of building reusable, flexible and scalable programs for infrastructure purposes, rather than a series of one-off programs,
- Addressing and alleviating data holders' concerns over data security, patient privacy and proprietary interests.

DISTRIBUTED DATA NETWORK INFRASTRUCTURE: THE COMMON DATA MODEL

The purpose of any common data model is to standardize the format and content of data, such that standardized applications, tools and methods can be applied to them. Figure 1 is a schematic intended to represent, at a high-level, the process for populating a common data model using healthcare claims data from a data partner site. In this example, the data partner site is a health insurance company or an integrated managed care consortium.

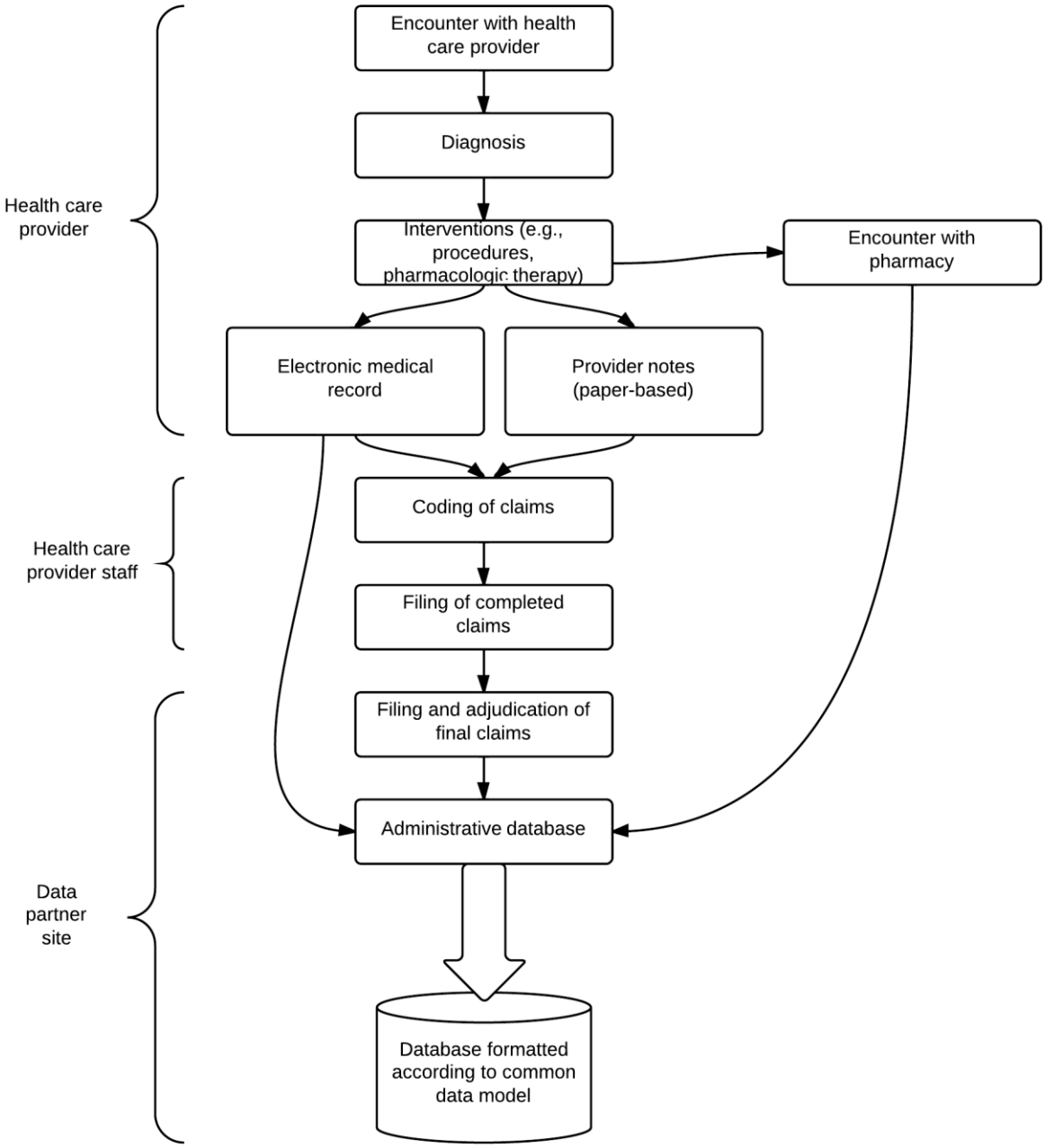

Figure 1. Process for populating a common data model using healthcare data from a data partner site

As patients interact with the healthcare delivery system, those interactions are captured in electronic medical record systems and/or administrative claims data systems. The data partner site in this example maintains its data in a central administrative database or warehouse and then converts those data into a common data model according to model specifications. All of these data reside behind the data partner's firewall, and that data partner maintains control over the usage and transfer of any of their data at all times.

DISTRIBUTED DATA NETWORKS IN EXISTANCE

The list below includes the names of healthcare-related distributed data networks in existence. This is by no means meant to be an exhaustive list but is rather intended to be used for illustrative purposes.

- FDA Mini-Sentinel
- PCORnet: The National Patient-Centered Clinical Research Network
- Innovation in Medical Evidence Development and Surveillance (IMEDS) Project.
- NIH Health Care Systems Research Collaboratory
- HMO Research Network
- Cancer and Cardiovascular Research Networks
- Vaccine Safety Datalink

Many of these networks have a particular focus. For example, the Mini-Sentinel project was funded to build an active, prospective surveillance system for health product safety; PCORnet focuses on conducting comparative effectiveness and patient-centered outcomes research; and the NIH Health Care Systems Research Collaboratory's focus is to improve the way clinical trials are conducted by creating infrastructure for collaborative research.

Several of these networks share the same data holders/partners. Some may also share the same common data model and, potentially, other infrastructure as the backbone to support the manner in which their network operates and analyzes data.

INTRODUCTION TO OUR CASE STUDY: MINI-SENTINEL

Mini-Sentinel is a pilot project sponsored by FDA to create a distributed data network and supporting infrastructure in order to enable an active surveillance system for monitoring the safety of drugs, biologics and devices—effectively, any FDA-regulated product—in the United States.

Section 905 of the Food and Drug Administration Amendments Act (FDAAA) of 2007 mandated the FDA to develop an enhanced ability to monitor the safety of drugs after they reach the market. This system, named Mini-Sentinel in its pilot phase (and to be called Sentinel thereafter), is intended to augment FDA's existing post-market safety monitoring systems. Current systems rely on FDA gathering information about their regulated products through programs that rely on external sources (product manufacturers, consumers, patients, and healthcare professionals) to report suspected adverse reactions to FDA. This type of safety monitoring is known as "passive surveillance." In contrast, Sentinel is intended to be an "active surveillance" system, as it will enable FDA to initiate its own safety evaluations that use available electronic healthcare data to investigate the safety of medical products.

The Mini-Sentinel Distributed Database (MSDD) currently consists of quality-checked data held by 18 partner organizations (health insurers or managed care consortiums). As of July 2014, the MSDD contained data on 178 million individuals, nearly 400 million person-years of observation time, 4 billion outpatient dispensings and 4 billion unique medical encounters.

Data partners standardize their data from their source systems into the Mini-Sentinel Common Data Model (MSCDM) and store those datasets as SAS datasets. Each site maintains physical control and ownership of their data, controls all uses of their data and controls all transfer of their data.

Figure 2 depicts the various tables included in the MSCDM. The MSCDM consists of a suite of several tables; six of the tables are considered 'core' and are present across all data partner sites (enrollment, demographic, outpatient dispensing, encounter, diagnosis, procedure). There are additional tables that are considered ancillary, as they are not present at all sites. Those include death, cause of death, laboratory tests and vital signs. The project is also actively exploring the integration of two additional inpatient data sources (inpatient dispensing and inpatient transfusions) into the common data model.

Data partners refresh their source data into MSCDM-formatted data quarterly to annually, depending on the site.

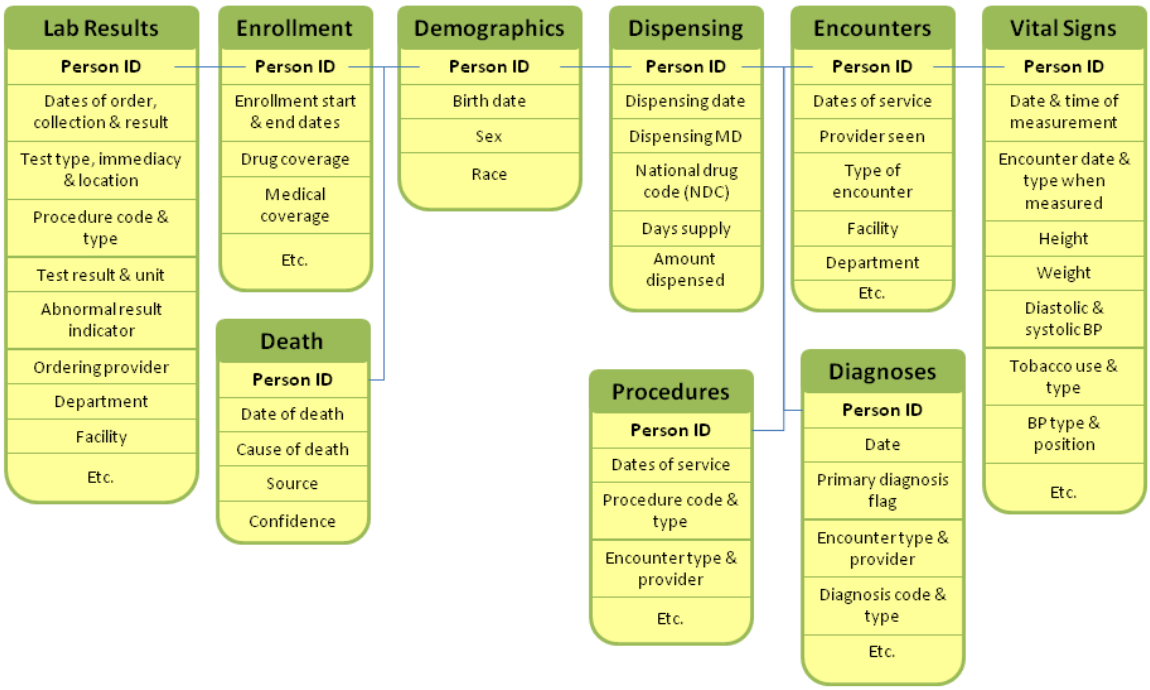

Figure 2. Mini-Sentinel Common Data Model

MINI-SENTINEL ANALYTIC FRAMEWORK

INTRODUCTION

The core concepts of building analytic programming infrastructure in a distributed data network are to recognize analytic- and programming-approach patterns where they exist, routinize programming tasks whenever possible, approach all programming tasks with reusability and flexibility in mind (rather than producing a series of one-offs), and to not reinvent the wheel.

The MSOC stresses the programming-related concepts of reusability and flexibility because there is so much overlap in the analytic approaches used across our surveillance and research projects. One project may be interested in studying a cohort of patients exposed to drug A that experienced outcome event X, while another project may be interested in a different exposure-outcome pairing but an otherwise similar analysis. Building flexible programs with regard to study parameters saves programming and auditing time and effort, and ensures consistency in analytic approaches across studies.

SAS-BASED, OPEN-SOURCE ANALYTIC PROGRAMMING TOOLS

The MSOC has developed a set of flexible, reusable SAS programs that were built to allow for rapid implementation of common epidemiologic and pharmacoepidemiologic study design methods. Each program is designed to execute against MSCDM-formatted data and can be customized using a wide variety of macro parameter settings and input files that define exposures, outcomes, covariates, inclusion/exclusion criteria, date ranges, age ranges, and other study protocol details.

These standard-approach, parameterized programs provide for considerable flexibility while significantly reducing programming time and subsequent program audit/review efforts. All MSOC-maintained core infrastructure SAS programs have been audited, pre-tested and validated. Similarly, all output produced is consistent and predictable across all data sites, stream-lining data aggregation and analytic reporting activities.

Collectively, as depicted in Figure 3, this suite of SAS programs comprises the Mini-Sentinel analytic framework. All programs described in this paper are available for download from the Mini-Sentinel public website and are accompanied by detailed documentation.

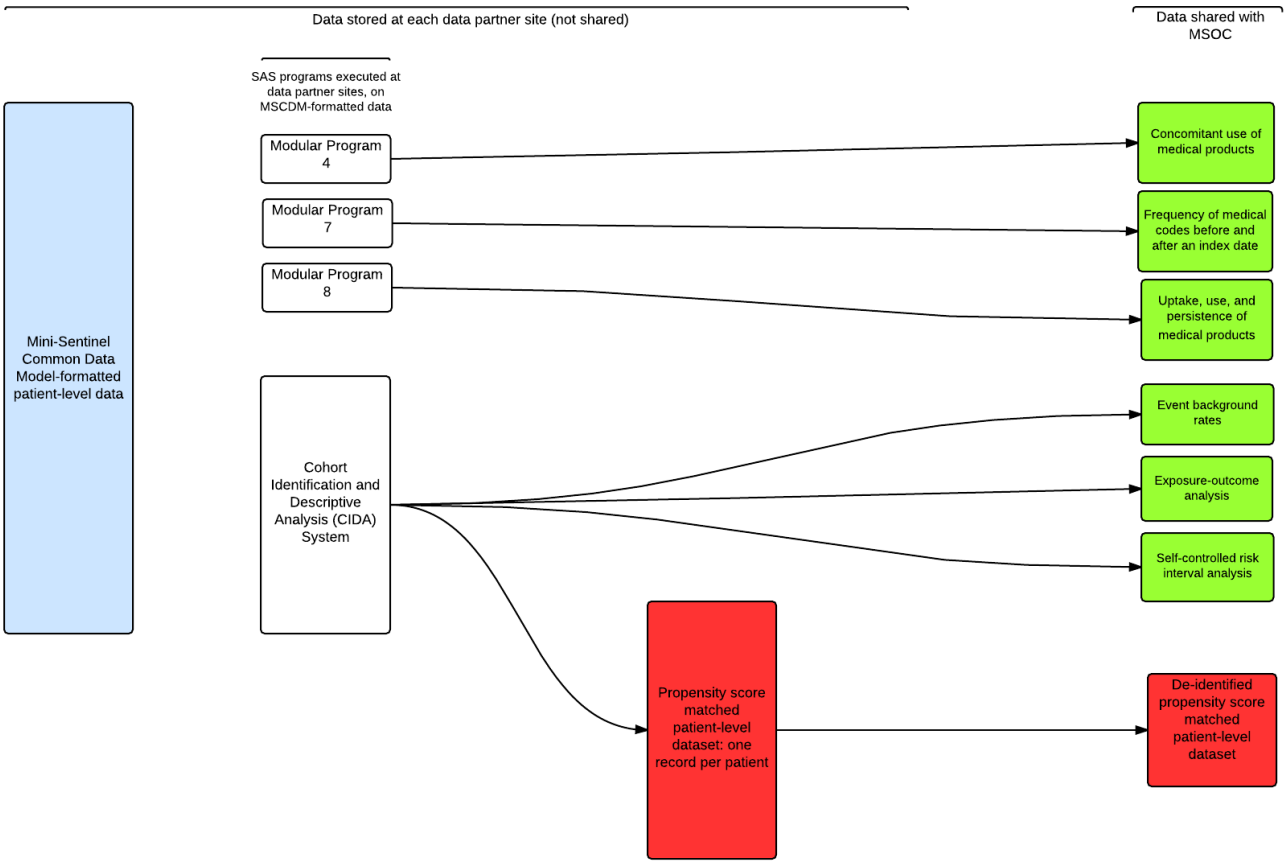

Figure 3. Schematic of the Mini-Sentinel analytic framework

"Modular" Programs

The MSOC developed and continues to maintain three "modular programs" that each perform a discrete descriptive analysis. Each can be used to characterize cohorts of interest and output various metric about those cohorts.

- Modular Program 4: identifies and characterizes concomitant use of medical products and occurrence of health outcomes of interest,
- Modular Program 7: identifies and characterizes the frequency of medical codes before and after an index date,
- Modular Program 8: identifies and characterizes the uptake, use, and persistence of medical products.

The analytic capabilities of each of these programs will be integrated into a future release of MSOC's principal analytic framework system, the Cohort Identification and Descriptive Analysis (CIDA) system. The remainder of this

paper will focus on the design and analytic capabilities of the CIDA system, as it serves as the backbone of the Mini-Sentinel analytic framework.

Cohort Identification and Descriptive Analysis (CIDA) System

The CIDA system has the capability to flexibly identify and extract cohorts of patients from raw MSCDM-formatted data based on a variety of user-specified cohort identification-related options (e.g., study dates, exposure definition, outcome definition, incidence criteria, inclusion/exclusion criteria, continuous enrollment requirements, relevant age groups, and so forth). Nearly all aspects relating to the identification of a study cohort have been built into the system as user-specified macro parameters or as user-supplied input files (formatted according to CIDA specifications). CIDA programming code has been constructed as a suite of over 30 highly-parameterized SAS macros, each performing a distinct function, making it a flexible, transparent and easily maintainable system.

There are three cohort identification strategies currently available within the CIDA system, each with a similar approach to the raw MSCDM-formatted data tables, but each generating different output metrics.

The first cohort identification strategy (the computation of event background rates) identifies an exposure, outcome, or medical condition within the MSDD and outputs metrics on the number of individuals with the exposure/outcome/medical condition, eligible members, and eligible member-days. Unadjusted rates are reported overall and stratified by user-defined age group, sex, year, and year-month. Figure 4 shows a schematic for the analytic approach taken to compute these background rates, while Figure 5 shows an example report that could be built from the output generated by this strategy.

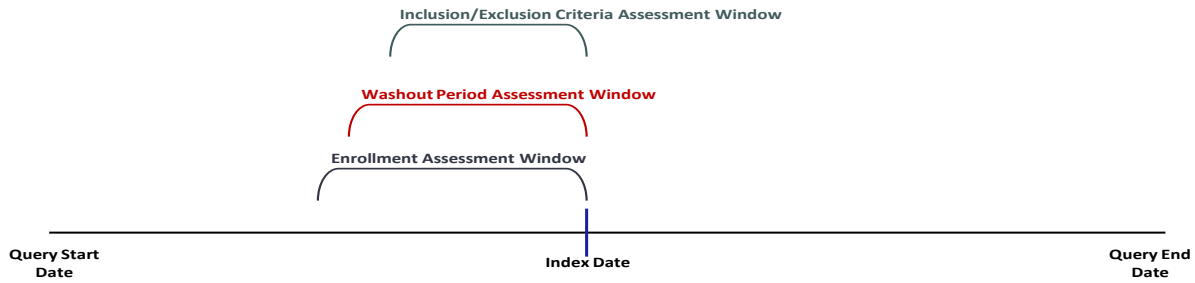

Figure 4. Schematic of analytic approach to computing background rates of events.

Below are some key highlights to the CIDA system's flexibility in identifying a study cohort:

- Study start and end dates are user-specified
- Ages or age groups for study inclusion are user-specified
- Index exposures or events can be defined using any combination of NDCs, diagnosis and procedure codes, and laboratory result values
- Continuous enrollment episodes are created using a user-specified enrollment gap and coverage type(s) (medical only, drug only, medical and drug). These continuous enrollment episodes are used to identify eligible members and calculate eligible member-days
- Incident use/events can be defined simply (e.g., no evidence of exposure in XX days before the index date) or using more complex criteria (e.g., no evidence of the exposure's drug ***class*** in XX days before the index date)

- Users can define any combination of inclusion and/or exclusion criteria using a number of days before or after index date
- Inclusion/Exclusion criteria are user-specified as any number of days before or after the index date, using any combination of NDCs, diagnosis and procedure codes, and laboratory result values
- Exposures or events at any point-of-view in a study (index, inclusion/exclusion, post-index events of interest, post-index censoring events) are user-specified using simple Boolean logic (a string of medical codes separated by “OR” logic), more complex Boolean logic (codes strung together by “AND” or “NOT”) and/or complex temporal relationships (two codes or a set of codes occurring with X days of each other)

Table 1: Summary of Medical Product "A" Use in the MSDD between January 1, 20XX and June 30, 20XX

	Users	Episodes	Dispensings	Days Supplied	Amount Supplied	Eligible Members	Member-Years	Users / 1K Eligible Members	Days Supplied/ User	Dispensings / User	Days Supplied/ Dispensing
Medical Product A Use, overall											
	2,235	2,499	2,499	2,499	2,499	118,455,039	338,206,231.9	0.02	1.12	1.12	1.00
Medical Product A Use, by age group											
0-9 Years	52	52	52	52	52	18,688,721	40,560,583.0	0.00	1.00	1.00	1.00
10-19 Years	39	67	67	67	67	20,092,977	44,827,760.6	0.00	1.72	1.72	1.00
20-29 Years	102	180	180	180	180	24,765,078	41,173,091.0	0.00	1.76	1.76	1.00
30-39 Years	279	300	300	300	300	22,775,195	44,789,320.5	0.01	1.08	1.08	1.00
40-49 Years	540	600	600	600	600	22,453,057	49,838,104.2	0.02	1.11	1.11	1.00
50-59 Years	320	359	359	359	359	19,315,738	46,510,593.4	0.02	1.12	1.12	1.00
60-69 Years	459	487	487	487	487	12,063,995	29,356,320.2	0.04	1.06	1.06	1.00
70-79 Years	234	244	244	244	244	5,607,625	16,007,536.8	0.04	1.04	1.04	1.00
80+ Years	210	210	210	210	210	3,256,894	10,142,865.3	0.06	1.00	1.00	1.00

Figure 5. Sample report based on output from the first cohort identification strategy

The second cohort identification strategy (an exposure-outcome analysis) builds on the logic introduced in the first strategy, by identifying an exposure of interest and determining medical product exposed time based on either a user-specified number of days after exposure initiation or on drug dispensing days of supply. The program then analyzes the data for the occurrence of health outcomes of interest during that exposed time. Output metrics include the number of exposure episodes and number of individuals exposed, number of health outcomes of interest, and days at-risk. Events per person-day at-risk are reported overall and stratified by user-defined age group, sex, year, and year-month. Unadjusted and adjusted (by age, sex, year and/or data partner site) incidence rate ratios (IRRs) can be calculated and compared between two identified cohorts (e.g., ratios between an exposed and a comparator cohort). Figure 6 shows a schematic for the analytic approach taken to compute these metrics for an exposure-outcome approach, while Figure 7 shows an example report that could be built from the output generated by this strategy.

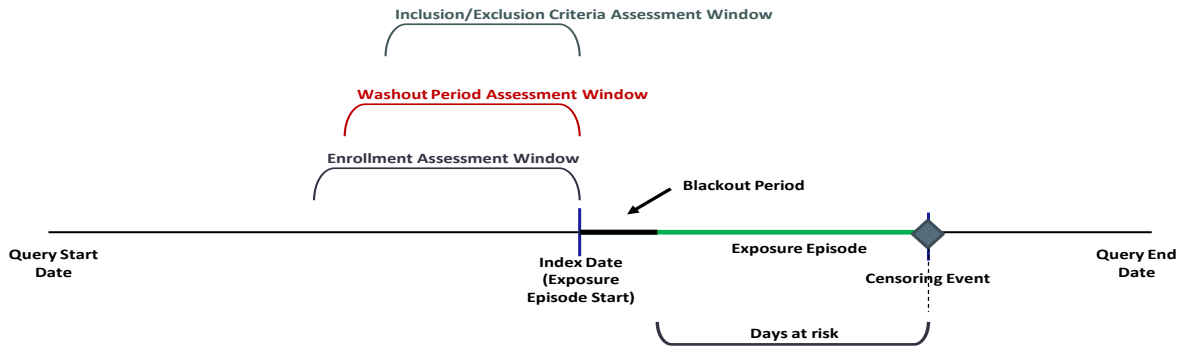

Figure 6. Schematic of the analytic approach to an exposure-outcome cohort identification strategy

Below are some key highlights of the ways the exposure-outcome cohort identification strategy builds on the prior-discussed strategy:

- Exposure episodes can be created in two ways:
    - Using the outpatient pharmacy dispensings' days of supply values to create continuous episode of treatment (using a stockpiling algorithm and a user-specified gap tolerance to handle overlaps and gaps in dispensing)
    - Using a user-defined duration of post-index exposure
- Users can specify episode extensions and minimum episode durations
- User-specified blackout periods may be supplied as the number of days post-exposure where a health outcome of interest will not be attributed to exposure. Exposure episodes with a health outcome of interest during the blackout period are discarded from analysis.
- Exposure episodes are censored at disenrollment/end of available data and at the occurrence of a user-specified health outcome of interest (defined by the user as any combination of NDCs, diagnosis/procedure codes, and lab result values). Users also have the ability to provide additional censoring criteria, including:
    - Any other user-specified censoring event (defined by the user as any combination of NDCs, diagnosis/procedure codes, and lab result values)
    - Evidence of death

Table 1: Summary of Medical Product A Use and Risk for Outcome Y in the MSDD between January 1, 20XX and December 31, 20XX

	New Users	New Episodes	Dispensings	Days Supplied	Amount Supplied	Years at Risk	Episodes w/ Events	Eligible Members	Member-Years	Percent of New Epsiodes with New Events	New Users / 1K Eligible Members	Days Supplied/ User
Medical Product A, overall	9,700	9,700	31,610	1,222,005	1,573,445	3,222.9	1,824	1,488,181	1,231,216.1	18.80	6.52	125.98
Medical Product A, by age group												
20-44 Years	233	233	715	24,860	34,063	61.8	45	97,224	56,494.8	19.31	2.40	106.70
45-64 Years	3,129	3,129	10,317	369,078	492,975	967.0	563	458,389	331,496.5	17.99	6.83	117.95
65-74 Years	2,695	2,695	8,689	348,946	441,165	928.2	514	390,814	288,731.9	19.07	6.90	129.48
75-84 Years	2,714	2,714	8,931	366,299	463,001	968.6	522	412,916	326,641.3	19.23	6.57	134.97
85+ Years	929	929	2,958	112,822	142,241	297.3	180	263,030	227,851.6	19.38	3.53	121.44

Figure 7. Sample report based on output from the exposure-outcome cohort identification strategy

The third cohort identification strategy (a self-controlled risk interval design) is a confounder-adjustment approach by design. This approach identifies an exposure of interest, identifies user-specified risk and control windows relative to the exposure date, and examines the occurrence of health outcomes of interest during the risk and control windows. Output metrics include the number of exposure episodes, exposed individuals, individuals with a health outcome of interest in the risk and/or control windows, and censored individuals, overall and stratified by user-defined age group, sex, year, year-month, and time-to-event in days. Figure 8 depicts a schematic for this cohort identification strategy.

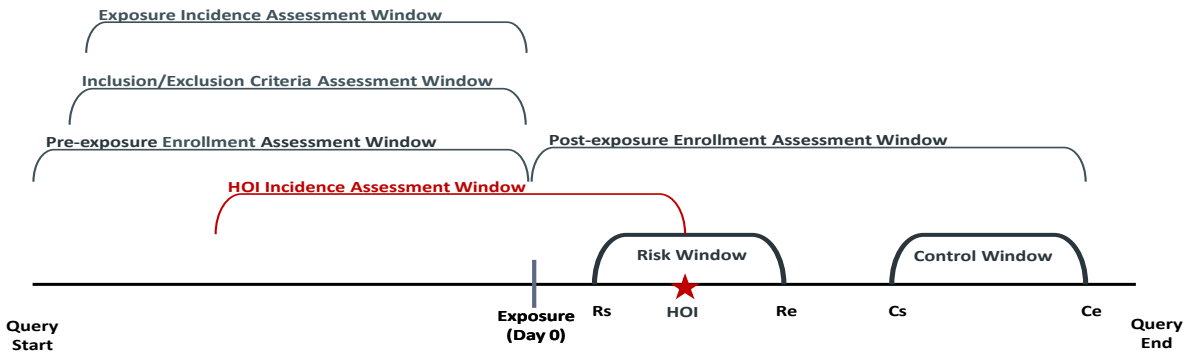

Figure 8. Schematic of the self-controlled risk interval cohort identification strategy

Below are some highlighted features of the self-controlled risk interval cohort identification strategy:

- Duration of the risk and control windows are user-specified, and the control window can occur pre- or post-exposure
- Post exposure enrollment is required for the self-controlled risk interval design. Patients must be enrolled from the exposure date to the end of the control window (or risk window, if the control window is before exposure).

- The analytic cohort identified using this method includes only informative patients (i.e., patients with a health outcome of interest in the risk or control window).
- Health outcome of interest incidence assessment is a user-specified number of days before the health outcome of interest date, not the index date.

Output metrics from the self-controlled risk interval cohort identification strategy include:

- Aggregated counts of patients exposed, exposure periods, events in risk and control windows, stratified by age group, sex, year, year-month, and time-to-event
- "Exposure cohort" data are retained (i.e., all patients meeting criteria before health outcomes of interest and post-exposure enrollment assessed)
- Allows characterization of patients who were censored due to disenrollment or death and number of patients with no events in the risk or control windows

Propensity score estimation and matching using the CIDA system

Given the observational nature of all of the data in the MSDD, unadjusted counts and rates of exposures and/or outcomes using the first or second cohort identification strategies are useful in a prep-to-research capacity but not very useful from a confounder-adjustment perspective. In order to address confounding in exposure-outcome types of analyses, the CIDA system includes the ability to estimate propensity scores using a set of user-specified covariates, with the option to compute and use a set of empirically-derived covariates (a high-dimensional propensity score approach), and to then match exposed-cohort patients to comparator-cohort patients on their estimated propensity scores based on a flexible set of user-specified criteria, including the caliper (0.01, 0.025 or 0.05) used in either a fixed 1:1 or a variable 10:1 ratio match strategy.

The standard output from a propensity score estimation and matching analysis includes measures of covariate balance, including absolute and standardized differences between unmatched and matched cohorts on the set of model covariates, as well as histograms depicting the propensity score distributions for each cohort. The output also includes the number of patients in each cohort group, the number matched from each cohort, the number that experienced the health outcome of interest, and the mean person-time of follow-up.

The dataset returned to the MSOC from this analysis is a de-identified, patient-level dataset containing:

- Demographic characteristics (sex, race, age at index)
- Matched-set identifiers
- Summarized baseline drug and medical utilization indicators or scores for each patient
- An indicator for whether the event of interest was experienced
- Follow-up time, in days
- Estimated propensity score

This dataset could be used for a variety of analyses. The CIDA system includes a program that enters these data into a Cox regression model that is stratified on the matched set in order to estimate hazard ratios and 95% confidence intervals.

Future propensity score matching-based enhancements using the CIDA system

A future propensity score matching-based approach within the CIDA system that is currently under development will include the ability to suppress the return of a de-identified patient-level dataset to the MSOC in favor of an aggregated dataset that will be summarized to the risk-set level. This approach will preclude the return of patient-level data altogether, while still being able to yield analogous effect estimates by way of a case-centered logistic regression model using the aggregated risk-set data (as opposed to a proportional hazards model using patient-level data). This enhancement will render the return of any patient-level data unnecessary, thus further minimizing data sharing while preserving the analytic capability to adjust for confounding.

CONCLUSIONS

Distributed data networks represent an emerging method of data sharing that increases data scope and coverage **and** addresses data security and patient privacy issues. Using the Mini-Sentinel project as a case-study, this paper demonstrates the enormous potential that distributed data networks hold for epidemiologic and

pharmacoepidemiologic studies by allowing access to greater volumes of data, without sacrificing privacy or analytic capabilities. This paper also demonstrates the contributions that distributed data networks have already made to the healthcare analytics community by investing in the building of data- and analytic programming-related infrastructure and by making those resources open-source and available to the public.

REFERENCES

- Bredfeldt CE, Butani A, Padmanabhan S, et al. 2013. “Managing protected health information in distributed research network environments: automated review to facilitate collaboration”. BMC Med Inform Decis Mak. 13:39.
- Brown JB and Platt R. 2013. “Distributed Research Networks: Lessons from the Field”. The Learning Health System Summit, Washington, DC. http://healthinformatics.umich.edu/sites/default/files/files/u24/7.Brown_.pdf . Accessed 20 October 2013.
- Curtis LH, Brown JB, Platt R. 2014. “Four Health Data Networks Illustrate The Potential For A Shared National Multipurpose Big-Data Network.” Health Affairs; 33(7): 1178-86.
- Food and Drug Administration Amendments Act of 2007, Pub. L. no. 110-85, Page 121 Stat. 944 (2007). http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/html/PLAW-110publ85.htm . Accessed 20 October 2013.
- Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. 2011. A combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol. 64(7):749-59.
- Hamilton J. 2012. “What Do You Mean, Not Everyone Is Like Me: Writing Programs For Others To Run”. Proceedings of the 2012 SAS Global Forum, Orlando, FL. http://support.sas.com/resources/papers/proceedings12/229-2012.pdf . Accessed 20 October 2013.
- Langston R. 2009. “Scalability of Table Lookup Techniques”. Proceedings of the 2009 SAS Global Forum, Washington, DC. http://support.sas.com/resources/papers/proceedings09/037-2009.pdf . Accessed 09 Jun 2014.
- “Modular Program 4: Frequency of Select Events During Concomitant Exposure to a Drug/Procedure Groups of Interest (version 5.0).” Mini-Sentinel Routine Querying Tools. June 11, 2014. Accessed April 2, 2015. Mini-Sentinel Routine Querying Tools. Retrieved April 2, 2015, from http://www.mini-sentinel.org/data_activities/modular_programs/details.aspx?ID=112.
- “Modular Program 7: Modular Program 7: Drug Use, Medical Diagnoses, and Medical Procedures Before and After an Exposure or Event of Interest (version 5.0).” Mini-Sentinel Routine Querying Tools. June 11, 2014. Accessed April 2, 2015. Mini-Sentinel Routine Querying Tools. Retrieved April 2, 2015, from http://www.mini-sentinel.org/data_activities/modular_programs/details.aspx?ID=138.
- “Modular Program 8: Modular Program 8: Uptake, Use, and Persistence of New Molecular Entities (NMEs) (version 1.0).” Mini-Sentinel Routine Querying Tools. June 27, 2014. Accessed April 2, 2015. Mini-Sentinel Routine Querying Tools. Retrieved April 2, 2015, from http://www.mini-sentinel.org/data_activities/modular_programs/details.aspx?ID=164.
- Nelson GS and Zhou J. 2012. “Good Programming Practices in Healthcare: Creating Robust Programs”. Proceedings of the 2012 SAS Global Forum, Orlando, FL. http://support.sas.com/resources/papers/proceedings12/417-2012.pdf . Accessed 03 September 2014.
- Popovic JR. 2014. “Programming in a Distributed Data Network Environment: A Perspective from the Mini-Sentinel Pilot Project.” Proceedings of the 2014 MidWest SAS Users Group Conference, Chicago, IL. http://www.mwsug.org/proceedings/2014/PH/MWSUG-2014-PH04.pdf. Accessed 01 Apr 2015.
- Psaty BM and Breckenridge AM. 2014. “Mini-Sentinel and Regulatory Science — Big Data Rendered Fit and Functional”. N Engl J Med; 370:2165-7.
- "Routine Querying System." Mini-Sentinel Routine Querying Tools. December 23, 2014. Accessed April 2, 2015. Mini-Sentinel Routine Querying Tools. Retrieved April 2, 2015, from http://www.mini-sentinel.org/data_activities/modular_programs/details.aspx?ID=166.
- Sherman T and Ringelberg A. 2013. “Defensive Programming and Error-handling: The Path Less Travelled”. Proceedings of the 2013 PharmaSUG Conference, Chicago, IL. http://www.pharmasug.org/proceedings/2013/TF/PharmaSUG-2013-TF24.pdf . Accessed 03 September 2014.
- Toh S, Reichman ME, Houstoun M, et al. 2013. “Multivariable confounding adjustment in distributed data networks without sharing of patient-level data”. Pharmacoepidemiology and Drug Safety; 22(11):1171-7.
- Toh S, Shetterly S, Powers JD, Arterburn D. 2014. “Privacy-preserving analytic methods for multisite comparative effectiveness and patient-centered outcomes research.” Med Care; 52(7):664-8.

ACKNOWLEDGMENTS

The author would like to extend a special thank you to all staff at the Mini-Sentinel Operations Center (MSOC). The expertise and ideas of my colleagues were invaluable to the development of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jennifer R. Popovic, DVM, MA

Harvard Pilgrim Health Care institute/Harvard Medical School

133 Brookline Ave, 6th Floor

Boston, MA 02215

617.509.9811

jennifer_popovic@harvardpilgrim.org