Lessons Learned from an EHR-Based Population Health Datamart: Southeastern Diabetes Initiative (SEDI)

Ursula A. Rogers[1], Shelley A. Rusincovitch, MMCi[1], Matthew Phelan, MS[2], N. Benjamin Neely, MS[3], Benjamin A. Goldstein, PhD[2]

[1]Duke Forge, Durham, NC; [2]Duke Clinical Research Institute, Durham, NC; [3]Duke Crucible, Durham, NC

Background

The **Southeastern Diabetes Initiative (SEDI)** was formed to improve population-level diabetes management, health outcomes, and quality of life for diagnosed and undiagnosed adults living with Type 2 diabetes mellitus.

The **SEDI datamart** was created in response to this initiative and includes data generated from routine patient care and delivery of health services as captured in the Electronic Health Record (EHR).

The datamart, evolving since 2012, is primarily a QI project with a secondary research focus, and is considered a great asset to researchers in the Duke community. This poster addresses **the pros and cons encountered when developing and utilizing such a population health datamart**.

Methods

The SEDI datamart has been **curated over a 5-year period**. A single protocol allows for the refresh of the datamart itself, and over fourteen (14) analysis protocols are have used the datamart to conduct specific research. To date, five (5) versions of the datamart have been generated, with each version adding new data or features.

Observation Period

- The most recent version of the SEDI datamart ("the SEDI 10-year dataset") contains EHR data collected from **1/1/2007 through 12/31/2016**

Population Definition

The cohort is defined as patients whom at some point during the observation period:

- **Lived in Durham County, NC**
- **had any type of encounter** with Duke Health and the Durham County Federally Qualified Health Center (FQHC), and
- were **18 years of age or older**

Data Domains

- The SEDI project integrates and harmonizes data from disparate sources and formats through use of a Common Data Model (CDM)
- The **PCORnet CDM**[2] was selected to represent patients and clinical data associated with their health system encounters, and was **extended to include additional SEDI-specific domains** permissible in the PCORnet specification (e.g. Address History)
- Ten (10) years of source data are transformed into the PCORnet model to generate the SEDI domains, as shown in Figure 1

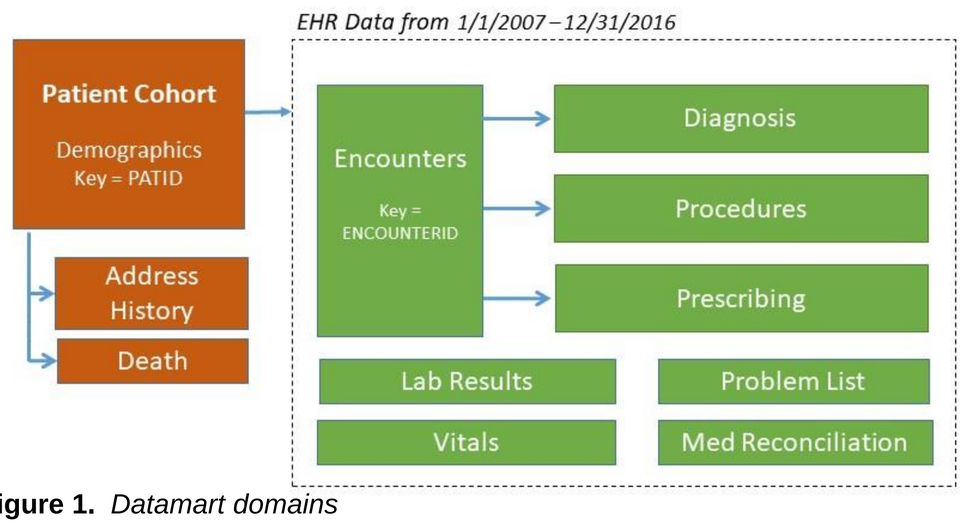

Figure 1. *Datamart domains*

With a **cohort of 385,202 patients**, the most recent version of the SEDI datamart contains a wealth of information which can be used to answer a diverse array of research questions. Up to 10 years of clinical data for these patients results in over 12 million visits, 42 million diagnosis, and 98 million lab tests.

# PATIENTS	385,202
# ENCOUNTERS	12,121,619
# DIAGNOSIS	42,924,315
# PROCEDURES	43,960,429
# PRESCRIBED MEDS	13,180,182
# LAB RESULTS	98,781,554
# MED RECONCILIATIONS	44,558,274
# VITALS	12,116,760
# PROBLEM LISTS	2,918,233
# DEATHS	25,063

Table 1. *10-yr SEDI datamart domain record counts*

Evaluation

As each version of the datamart was built, and as each customer using the datamart conducted their research, **we identified a list of "lessons learned" informing us of both strengths and limitations in our design choices**. As such, the following points should be considered when building a similar population health datamart:

Strengths	*Considerations*
Countywide cohort contains both healthy and unhealthy patients, is not restricted to Diabetes and is valuable to a breadth of use cases including research, quality improvement, and operations	Broad research potential
Contains longitudinal representation of a patient's interaction with the Duke Health/FQHC systems	Merges legacy and current EPIC system data; tells the story of a patient's health journey
With each refresh, data domains were added or expanded onto	Flexible model, yet predictable base domains
Streamlined regulatory structuring with a single protocol for the datamart itself, and separate analysis protocols for each research project	Facilitates "easy" research on data already curated; IRB efficient
Built in same Oracle schema as Epic and the data warehouse[3] for ease of technical implementation	Infrastructure is efficient for refreshes
Big data focus lends itself well to machine learning projects	Big data is rich for predictive modeling
Domain curation logic and transformation in the codebase have been adapted for many other EHR-based projects	Code re-use
Limitations	
Data is restricted to those patients residing in one county (Durham, NC) who were at least 18 years of age	Inherently limiting; Missing pediatric and a significantly-ill population traveling from out-of-state
Difficult to merge different source systems into PCORnet CDM (example: encounter definitions vary widely system to system)	Time-consuming and difficult mapping source data to CDM
Many RAW fields were added to the PCORnet CDM to accommodate local research IRB protocols	The PCORnet CDM did not fully meet SEDI's needs; perhaps there is a better fit
Some CDM domain structures did not meet research needs	The Vitals table is not in tabular form and must be deconstructed
Local data warehouse, Clarity and CDM changes occur often which require maintenance	Datamart refreshes often required significant re-work
For some use cases, too much data can be cumbersome to work with	Big data is difficult to work with if there are computing resource restrictions (e.g., Lab Results are over 100GB and difficult for some programs to process)

Table 2. *Considerations when building a population datamart*

Conclusion

The SEDI Datamart has been a valuable resource for the research community, providing a quality longitudinal look at a patient's clinical history. Current studies using this datamart involve studying countywide diabetes care, cardiovascular outcomes, medication adherence, socio-economic factors, diabetes phenotypes and predicting disease progression.

While the value in such a datamart is obvious, considerations (Table 2) must be taken into account when building such a large population dataset. In summary,

- ✓ **Population restrictions can have significant repercussions on potential future analyses**
- ✓ **Use of a Common Data Model should be carefully considered to ensure it is not too limiting, yet is flexible enough to be extended**
- ✓ **The size and breadth of the data should be carefully defined to appropriately meet research, quality improvement, and operational needs**

References / Acknowledgments

1. The projects and the work described in this poster are supported in part by Grant Number 1C1CMS331018-01-00 the Department of Health and Human Services, Centers for Medicare & Medicaid Services, and in part by the Bristol Myers Squibb Foundation Together on Diabetes program. The project is additionally supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant Award Number UL1TR001117 at Duke University. The contents of this poster are solely the responsibility of the authors and have not been approved by the Department of Health and Human Services, Centers for Medicare & Medicaid Services or the NIH.
2. PCORNET, The National Patient-Centered Clinical Research Network. Available at: https://pcornet.org/pcornet-common-data-model / [Accessed October 2018].
3. Additional thanks to the Data Warehouse Foundation team (Duke Health Technology Solutions) led by Michael Santoianni.

Contact Information: Ursula A. Rogers, Senior Informaticist
ursula.rogers@duke.edu