12-2017

Impact of Terminology Mapping on Population Health Cohorts IMPaCt

Barbara A. Berkovich
Barbara.A.Berkovich@uth.tmc.edu

Impact of Terminology Mapping on Population Health Cohorts
IMPaCt

By

Barbara Berkovich, M.A.

APPROVED:

_____
Susan Fenton, PhD, Chair

_____
Amy M Sitapati, MD

_____
Amy Franklin, PhD

Date approved: _____

Impact of Terminology Mapping on Population Health Cohorts - IMPaCt

A
Dissertation

Presented to the Faculty of
the University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Barbara Berkovich, MA

University of Texas Health Science Center at Houston

2017

Dissertation Committee:

Susan Fenton, PhD[1], Chair
Amy M Sitapati, MD[2]
Amy Franklin, PhD[1]

[1]The School of Biomedical Informatics
[2]UC San Diego School of Medicine

Dedication

This work is dedicated to my parents, Victor and Marilyn Kaufmann, and my husband Gil Berkovich for his incredible support.

I also dedicate this study to Larry Weed, inventor of the problem-oriented medical record methodology. His approach continues to resonate today.

> *"If you can't evaluate what you're doing, then there's a very serious possibility that you do not know what you're doing."*
>
> Larry Weed
> Physician, educator, inventor
> (1923-2017)

Acknowledgements

I would like to acknowledge my doctoral committee: Susan Fenton, PhD (Chair), Amy Sitapati, MD, and Amy Franklin, PhD for their guidance during my doctoral program and research.

To the faculty and staff at the University of Texas Health Science Center at Houston School of Biomedical Informatics, thanks to all who supported and facilitated this journey. I would especially thank Dean Sittig, PhD, Trevor Cohen, PhD, and Craig Johnson, PhD, for their instruction and guidance; Jamie Hargrave and the staff in the Office of Academic Affairs for providing so much information and support; Marcos Hernandez, for facilitating learning at a distance; and Mary Wickline for assistance with the literature search. The inspiration for this journey can be credited to Dean Jiajie Zhang's American Medical Informatics Association (AMIA) 10 x 10 course, Healthcare Interface Design and Distance Learning Course. Financial support for this course was provided by the UC San Diego Staff Equal Opportunity Enrichment Program (Summer 2012).

Abstract

Background and Objectives: The population health care delivery model uses phenotype algorithms in the electronic health record (EHR) system to identify patient cohorts targeted for clinical interventions such as laboratory tests, and procedures. The standard terminology used to identify disease cohorts may contribute to significant variation in error rates for patient inclusion or exclusion. The United States requires EHR systems to support two diagnosis terminologies, the International Classification of Disease (ICD) and the Systematized Nomenclature of Medicine (SNOMED). Terminology mapping enables the retrieval of diagnosis data using either terminology. There are no standards of practice by which to evaluate and report the operational characteristics of ICD and SNOMED value sets used to select patient groups for population health interventions. Establishing a best practice for terminology selection is a step forward in ensuring that the right patients receive the right intervention at the right time. **The research question is, "How does the diagnosis retrieval terminology (ICD vs SNOMED) and terminology map maintenance impact population health cohorts?"** Aim 1 and 2 explore this question, and Aim 3 informs practice and policy for population health programs.

Methods

Aim 1: Quantify impact of terminology choice (ICD vs SNOMED)

ICD and SNOMED phenotype algorithms for diabetes, chronic kidney disease (CKD), and heart failure were developed using matched sets of codes from the Value Set

Authority Center. The performance of the diagnosis-only phenotypes was compared to published reference standard that included diagnosis codes, laboratory results, procedures, and medications.

Aim 2: Measure terminology maintenance impact on SNOMED cohorts

For each disease state, the performance of a single SNOMED algorithm before and after terminology updates was evaluated in comparison to a reference standard to identify and quantify cohort changes introduced by terminology maintenance.

Aim 3: Recommend methods for improving population health interventions

The socio-technical model for studying health information technology was used to inform best practice for the use of population health interventions.

Results

Aim 1: ICD-10 value sets had better sensitivity than SNOMED for diabetes (.829, .662) and CKD (.242, .225) (N=201,713, p <= .001). ICD-10 had worse specificity than SNOMED for diabetes (.972, .975), but the same for CKD (p <= .001). Heart failure cohorts had no significant differences between ICD and SNOMED.

Aim 2: Following terminology maintenance the SNOMED algorithm for diabetes increased in sensitivity from (.662 to .683 (p <=0.001)). No change was observed in the performance of CKD and heart failure algorithms. Those cohorts were unaffected.

Aim 3: Based on observed social and technical challenges to population health programs, including and in addition to the development and measurement of phenotypes, a practical method was proposed for population health intervention development and reporting.

Discussion

As a measure of overall performance, the F score for ICD phenotypes for diabetes, CKD and heart failure equal to or better than SNOMED. Standardized testing and reporting practices for population health algorithms will inform local and national practice for management of population health cohorts.

Vita

1985.....B.S., Industrial and Systems Engineering, University of Southern California

2010.....M.A., Educational Technology, San Diego State University

2014 to present.....PhD Student, Health Informatics University of Texas Health Science Center at Houston, School of Biomedical Informatics

Publications

Sitapati, A., Kim, H., Berkovich, B., Marmor, R., Singh, S., El-Kareh, R., Clay, B., Ohno-Machado, L. (2017). Integrated precision medicine: the role of electronic health records in delivering personalized treatment. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 9(3).

Field of Study

Health Informatics

Table of Contents

List of Tables

List of Figures

List of Equations

Key Terms

Coding System	Any terminological system that uses codes for designating concepts. (de Keizer, Abu-Hanna, & Zwetsloot-Schonk, 2000)
Cohort	Any designated group followed or traced over a period of time. (cohort)
Classification	A classification uses more general "is-member-of" relationship. (de Keizer et al., 2000)
Gold standard	Gold standard refers to the method by which a reference standard is generated by two or more independent reviewers with adjudication to obtain agreement.
High-throughput clinical phenotyping	High-throughput clinical phenotyping executes an algorithm against already existing data within an EHR system to rapidly obtain a large pool of eligible study subjects. (Wei et al., 2012)
Phenotyping	Phenotyping is the action of applying an algorithm to select a cohort within an EHR system for a defined purpose, including case–control cohorts for genome-wide association studies, clinical trials, quality metrics, and clinical decision support. (Pathak et al., 2013)
Population health	Use of the EHR system to identify patient cohorts in need of evidence-based interventions, and to facilitate action to address care gaps.
Population health registry	An EHR-based registry for the purpose of driving clinical interventions is also called a population health registry. This registry subtype identifies care gaps and triggers bulk ordering and secure bulk messaging as well as clinic outreach phone calls. (Berkovich, 2016); (Sitapati, 2016)
Reference standard	A reference standard is the list of positive and negative case findings against which the performance of a phenotype algorithm is evaluated. Manual generation of a reference standard can be grouped into three levels: Gold standard, Trained standard and Regular practice. (Stanfill, Williams, Fenton, Jenders, & Hersh, 2010)

	Automated forms of developing reference standards show promise. (Agarwal et al. (2016)
Terminology	A list of terms related to concepts is a "terminology". In this sense, both ICD and SNOMED CT® are terminologies in the domain of clinical findings and diagnoses.
Thesaurus	An ordered terminology that includes synonyms (de Keizer et al., 2000)
Value Set	Numerical values (codes) and human-readable names (terms), drawn from standard vocabularies such as SNOMED CT® , RxNorm, LOINC and ICD-10-CM, which are used to define clinical concepts. For example, a value set may contain any number of codes across terminologies that represent a clinical concept such as a patient with myocardial infarction. These clinical concepts can then be used in constructing algorithms for quality measures, or population health identification rules. (U.S. National Library of Medicine)
Vocabulary	Terminology or Thesaurus that includes concept definitions. (de Keizer et al., 2000)

Chapter 1: Introduction

The goal of population health programs for chronic care is to improve the quality and while reducing unnecessary utilization such as emergency visits and hospitalizations (Altavela, Dorward, Sorrento, Diehl, & Wyman, 2017). Chronic disease management aims to avert morbidity and prioritize clinical interventions that can reduce risk for poor outcomes (McClatchey, 2001). Population health registries are a pivotal tool to support the delivery of this kind of evidence-based care (Lyon & Slawson, 2011). The clinicians using population health registries depend on accurate groups, or cohorts of individuals who share a characteristic and who are then followed forward in time. Grouping patients, with diabetes for example, is a phenotyping task which is conceptualized in Figure 1.

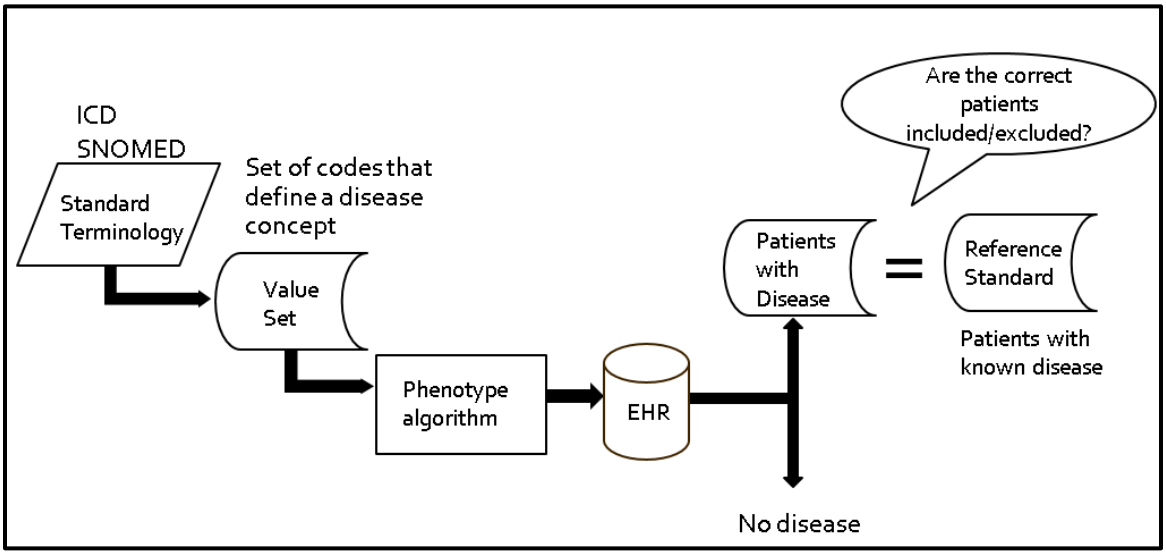


Figure 1. Disease phenotyping with standard terminologies

Diagnostic terms and their associated codes are applied to the patient records in the problem list, encounter diagnoses, and other locations within the EHR system depending on the nature and setting of the patient care. A phenotyping algorithm uses diagnosis codes to separate patients into cohorts, those with the specific code(s) associated with them. The algorithm matches patient diagnosis codes from the EHR data to a value set of codes that define a disease concept in a standard medical terminology. The International Classification of Diseases (ICD) and the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) are two of these terminologies. Both have mandates requiring their support in certified EHR systems. When a cohort is selected from a pool of patients, one way of measuring the accuracy of the assignment to a disease group is by comparing each patient in the cohort to a reference standard of patients known to have the disease.

Terminology maps allow patients to be identified for diagnosis groups using either ICD or SNOMED values sets. As terminologies evolve, the maps must be maintained to incorporate new codes or remove codes that are no longer in use.

The research question is, "How does the value set terminology (ICD vs SNOMED) and terminology map maintenance impact population health registries?"

Specific Aim 1 Quantify impact of terminology choice (ICD vs SNOMED)

Specific Aim 2 Measure the impact of terminology maintenance on SNOMED cohorts

Specific Aim 3 Recommend methods for improving population health interventions

This research will evaluate key risks related to terminology mapping as well as possible approaches to improve the visibility of terminology shifts that occur over time and mapping changes in the EHR that apply those terminology changes to the system.

1.1 Population Health in Practice

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 gave legislative authority to CMS to develop quality measures to help drive the development and the adoption of health information technologies. Since then, the number of incentivized quality measures at local, state and national levels has continued to increase. In addition to the Merit-based Incentive Payment System (MIPS) and Advanced Alternative Payment Models (APMs), CMS 1115(a) demonstration waivers are designating billions of dollars to states for Medicaid reform, a large portion of which will be tied to quality indicators. California's five-year Medi-Cal 2020 Demonstration program includes $6.2 billion of initial federal funding to transform and improve the quality of care, access, and efficiency of health care services for 12.8 million member (Mcleod, 2016). Roughly half of this sum is allocated to Public Hospital Redesign and Incentives in Medi-Cal (PRIME) which offers incentives public hospitals for performance measures for quality and efficiency. As value-based care models create a financial imperative in many organizations to qualify for government quality incentives, and benefit from shared risk programs, pressure to "meet the measures" is intensifying. In this environment, population health has become a preferred care delivery paradigm to drive health care quality and quality measures.

The Agency for Healthcare Research and Quality (AHRQ) National Measures Clearinghouse maintains a broad set of quality measures sponsored by diverse organizations that cover many disease states. Healthcare providers demonstrate their adherence to evidence-based guidelines through a suite of quality measures. Table 1 provides examples of Meaningful Use quality measures for diabetes (*eCQM 122-Percent of patients with hemoglobin A1C (HbA1C) levels > 9%*, and *134- Percent of patients with a screening for nephropathy*), End Stage Renal Disease (*Percent of patients with mean hemoglobin value greater than 12 g/dL), and* heart failure (*eCQM 135-Percent of Patients with heart failure (Left ventricular ejection fraction LVEF < 40%) who were prescribed ACE inhibitor or ARB therapy*).

Population health tools embedded within the EHR continue to be developed and refined to ensure that evidence-based interventions for chronic care management are routinely and reliably provided to patients as demonstrated by quality performance measures. Errors in the patient groups may negatively impact the quality and efficiency of patient care. The risk and cost of discovery errors (false positives) and omission errors (false negatives) are determined by the type of intervention and the number of patients in the population served. Table 1 also provides examples of the possible impact of discovery and omission errors. For patients in chronic care management programs, discovery errors may result in unnecessary messages and lab orders, wasted outreach costs, and possible drug-drug interactions. For these same groups of patients, omission errors may increase the risk and cost of disease complications, and put patients at higher risk of death. Misattribution of patients to population health cohorts may lower performance on quality measures as well, because the indicated interventions are not appropriate.

Table 1. *Meaningful Use quality measures for diabetes, end-stage renal disease and heart failure*

Meaningful Use Quality Measures		Discovery Error (False Positive)	Omission Error (False Negative)
Diabetes	HbA1c level > 9.0%	Unnecessary messages and lab orders	Increased risk/cost of complications
End-stage renal disease	Mean hemoglobin value greater than 12g/dL	Wasted outreach	Miss patients and put at higher risk of death
Heart failure	Heart failure (LVEF < 40%) who were prescribed ACE inhibitor or ARB therapy	Possible adverse drug reactions	Increased risk of death or re-admission

The Centers for Disease Control and Prevention (2017) estimates 23 million adults in the U.S. have a diagnosis of diabetes. Small errors in discovery or omission of patients in a chronic disease cohort are magnified when applied across a large population. In quality reporting a 1% error rate may be well within acceptable limits. However, when applied to actual clinical interventions, a false omission rate of 1% in a population of 23 million could put over 200,000 patients at risk for missing diabetes follow-up care. At that same rate of error, false discovery may cause another 200,000 to be targeted for diabetes interventions that were not appropriate.

The cost of treatment increases dramatically as patients move from a single disease state to multiple chronic diseases. The United States Renal Data System (2016) reports a 2014 fee-for-service expenditure of $254.4 billion on Medicare beneficiaries age 65 and older with diabetes, CKD and/or heart failure. The breakdown by disease category in Table 2 identifies 5.9 million beneficiaries with diabetes (with or without CKD and/or heart failure) comprising 24.02% of the Medicare population and utilizing 35.12% of the costs. The per person per year (PPPY) cost for diabetes alone ($12,116) increases significantly when diabetes is treated with CKD and/or heart disease ($16,003). As a percentage of the beneficiary population, CKD (10.77%) and heart failure (8.89%) utilize double the percentage of total budget than their numbers would suggest (CKD 20.77%, heart failure 20.60%). The goal of population health programs is to identify and manage chronic disease to keep patients healthier and reduce the burden of these diseases on the health system.

Table 2. *Medicare spending for beneficiaries aged 65 and older 2014*

	Count	Medicare Population %	Cost %	PPPY Cost 1 Condition	PPPY Cost 2 or more Conditions	Total Costs (millions)
All Diabetes	5.9 m	24.02	35.12	$12,116	$16,003	$89,327
All CKD	2.6 m	10.77	20.77	$15,673	$21,857	$52,819
All Heart failure	2.1 m	8.89	20.60	$20,733	$26,975	$52,409

Note. The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy or interpretation of the U.S. government.

1.2 Population Health Registries

Within a health system, population health programs utilize registries to ensure that patients sharing a common chronic disease receive the standard of care defined in an evidence-based protocol (Drawz et al., 2015). Regardless of etiology, chronic care management programs face the common challenge of monitoring a cohort of patients to ensure they are reevaluated at regular intervals, and that treatments are effective based on quantitative measures. The use of registries enables the identification of a group of patients at risk for multiple adverse outcomes, and creates the opportunity for efficient and directed intervention when there is a care gap between the patient's current treatment state and the protocol. A population health registry integrally embedded in an EHR system enables bulk orders, and secure bulk messaging via the EHR patient portal (Berkovich, 2016; Sitapati, 2016). Due to the very real impact on human resources, patients, and payers, inclusion in these registries requires the same high level of accuracy that has typically been associated with clinical decision support.

1.3 History of ICD and SNOMED

Both ICD and SNOMED CT® are standard medical terminologies in the domain of clinical findings and diagnoses. They support the input and retrieval of information from a clinical system (Brown & Sonksen, 2000). ICD codes have long been used by CMS to process claims for hospitals, clinics, and other professional services. In 2016, $366 billion in healthcare reimbursements was processed using ICD-10-CM codes to define the diagnoses covered by the Medicare Fee for Service Program (Centers for Medicare &

Medicaid Services (CMS), 2016). The Code of Federal Regulations 45 CFR 162.1002 designates ICD-10-CM as the medical data code set for medical problems for the period on or after October 1, 2015 (Code of Federal Regulations (annual edition), 2015). The Department of Health and Human Services maintains and distributes ICD-10-CM for the following conditions: (i) Diseases, (ii) Injuries; (iii) Impairments; (iv) Other health problems and their manifestations. (v) Causes of injury, disease, impairment, or other health problems (Code of Federal Regulations (annual edition), 2015).

Bowman (2005), Director of Coding Policy and Compliance for the American Health Information Management Association (AHIMA), described ICD as a "Classification system" that functions optimally for aggregating patient groups for claims processing and quality programs outputs. Classification systems were not intended or designed as the primary documentation for clinical care, yet they are the most common source of clinical data today, due to their necessity for claims processing (Bowman, 2005). The terminology has developed into an international standard for diagnostic classification in epidemiology, health management and clinical purposes (Fung & Xu, 2012).

From its inception, the necessity of revision based on scientific discovery was recognized, with the original ICD update cycle set at 10 years. ICD-9 was introduced to the public domain in 1977, formatted with 4-digit categories and optional 5 digit subdivisions (Hirsch et al., 2016). ICD-9-Clinical Modification (CM) was the U.S. extension developed by the National Center for Health Statistics to support diagnostic coding in the inpatient and outpatient settings (Hirsch et al., 2016). ICD-10 was first published in 1992 (World Health Organization). Figure 2 illustrates the differences between ICD-9 and 10 code formats.

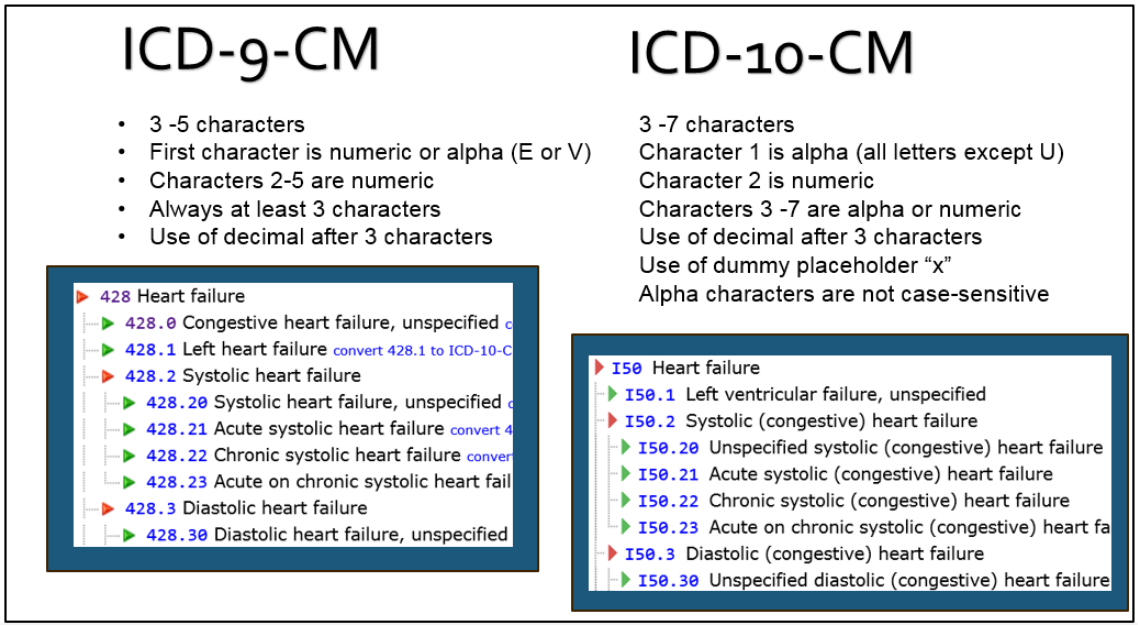


Figure 2. Sample of ICD-9-CM and ICD-10-CM billing codes

SNOMED CT® is also legislated for use in U.S. EHR systems as a vocabulary standard for representing electronic health information about medical problems per 45 CFR 170.207 (Code of Federal Regulations (annual edition), 2015). The International Health Terminology Standards Development Organization (IHTSDO) maintains and distributes SNOMED CT® "to facilitate the accurate recording and sharing of clinical and related health information and the semantic interoperability of health records" (Randorff Hojen & Kuropatwa, 2014). SNOMED CT® is considered the most comprehensive, multilingual clinical terminology in the world (Fung & Xu, 2012). It has a polyhierarchical logic model, and serves as a thesaurus, nomenclature, taxonomy, ontology, and coding system of clinical concepts (Saitwal et al., 2012). Despite these advantages for clinical documentation, SNOMED's size, complex hierarchies and lack of reporting rules render it impractical for patient reimbursement and regulatory reporting

(Bowman, 2005). A survey of EHR vendors by Giannangelo and Fenton (2008) found a lack of incentives or drivers in the industry was a barrier to SNOMED CT implementation within their products. Subsequent legislation mandating the use of SNOMED CT in EHR systems, and the selection of SNOMED as the reporting terminology for the CMS Merit-based Incentive Payment System (MIPS) have introduced these drivers and incentives into the industry. A 2013 article in the Journal of Biomedical Informatics finds that although it is reported to be used in over 50 countries, there is still much work ahead to bring SNOMED CT into routine clinical use (Lee, Cornet, Lau, & De Keizer, 2013). Quality challenges reported by this study included content coverage, hierarchical relationships, ambiguity of terms and syntactic consistency.

Bowman (2005) describes SNOMED-CT® as a "Reference terminology" designed to codify clinical information captured in an EHR during the course of patient care. As an input terminology, the semantic and contextual meanings of the clinical terms are paramount. The conceptual diagram of a SNOMED concept browser in Figure 3 illustrates that concepts exist within a hierarchy of parent-child relationships. For example, SNOMED concept 84114007 Heart Failure has a parent concept called Disorder of cardiac function and 26 child concepts including Acute Heart Failure 56675007, Cardiorenal syndrome 445236007, and Chronic heart failure 48447003. SNOMED CT concept hierarchies support retrieval of diagnosis data at various levels of granularity, and one SNOMED concept may be mapped to just one, or potentially thousands of ICD codes.

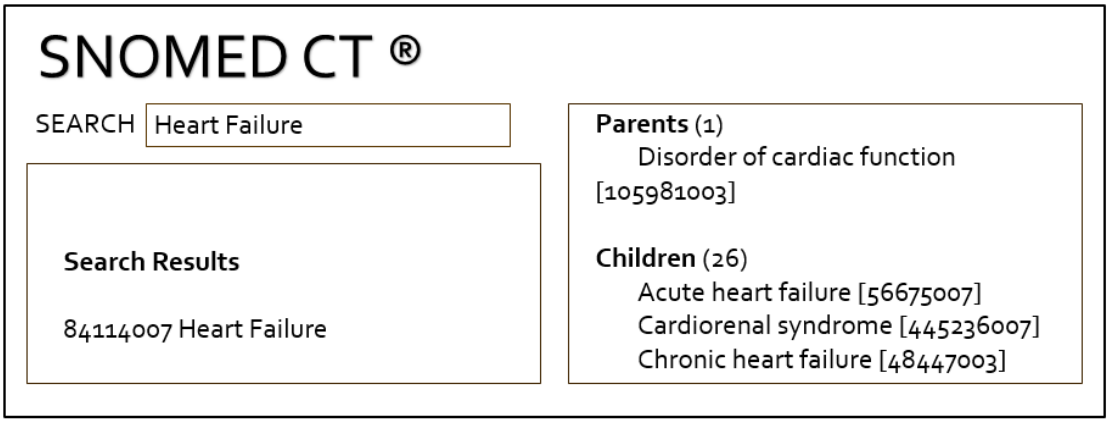


Figure 3. Conceptual diagram of a SNOMED CT® concept browser

The Journal of the American Medical Informatics Association (JAMIA) reported how three professional coding companies applied SNOMED concepts to the same clinical findings (Andrews, Richesson, & Krischer, 2007). In this small study, SNOMED codes for a vascular exam were selected to describe the finding, body structure and a qualifier. No significant correlation was found between the assigned codes. In fact, all three agreed on the core concept only 33% of the time, and 23% of the time there was no agreement at all, Andrews et al. (2007) raise the question, if coders can't agree on how to code a finding, what impact will that have on the retrieval?

1.4 Cross-terminology Mapping

Cross-terminology mapping in the EHR is essential because there is no single diagnosis terminology standard (Foley, Hall, Perron, & D Andrea, 2007). A 2005 White Paper by the American Health Information Management Association (AHIMA) described how ICD and SNOMED should be used together with mapping linking the two (Bowman, 2005). Figure 4 is an adaptation of the terminology mapping concept by which the SNOMED reference terminology is the input for clinical terms, and ICD is the output

classification system used for claims processing and quality reporting (Bowman, 2005). A terminology map is sometimes referred to as a crosswalk, implying a 1 to 1 relationship. In fact, because terminologies have different structures and intended uses, relationships can be: 1 to many; many to 1; many to many; or complex. Hussain et al. (2014) describe how erroneous mapping can be unintentionally created when two unrelated concepts (A1 and A2) are mapped to common concept (B). Although this logic applies in algebraic relationships (If A1 = B and A2 = B, then A1 = A2), this construct does not always hold up when applied to hierarchical terminological constructs. It is also incorrect to assume that if A1 equals B that all dependents of A1 in a hierarchical concept tree also equal B. Reich, Ryan, Stang, and Rocca (2012) describe cases where no mapping is possible such as when a code does not have a corresponding code in the destination terminology. The size and structure of the ICD and SNOMED terminologies present significant challenges to those creating and maintaining terminology maps (Boyd et al., 2015).

In the era of paper medical records, a workforce of coders was responsible for applying billing codes based on free text evidence in the health record. In EHR systems of today clinicians routinely perform a text search that returns a list of possible diagnosis terms from which one is selected to be applied to the patient record.

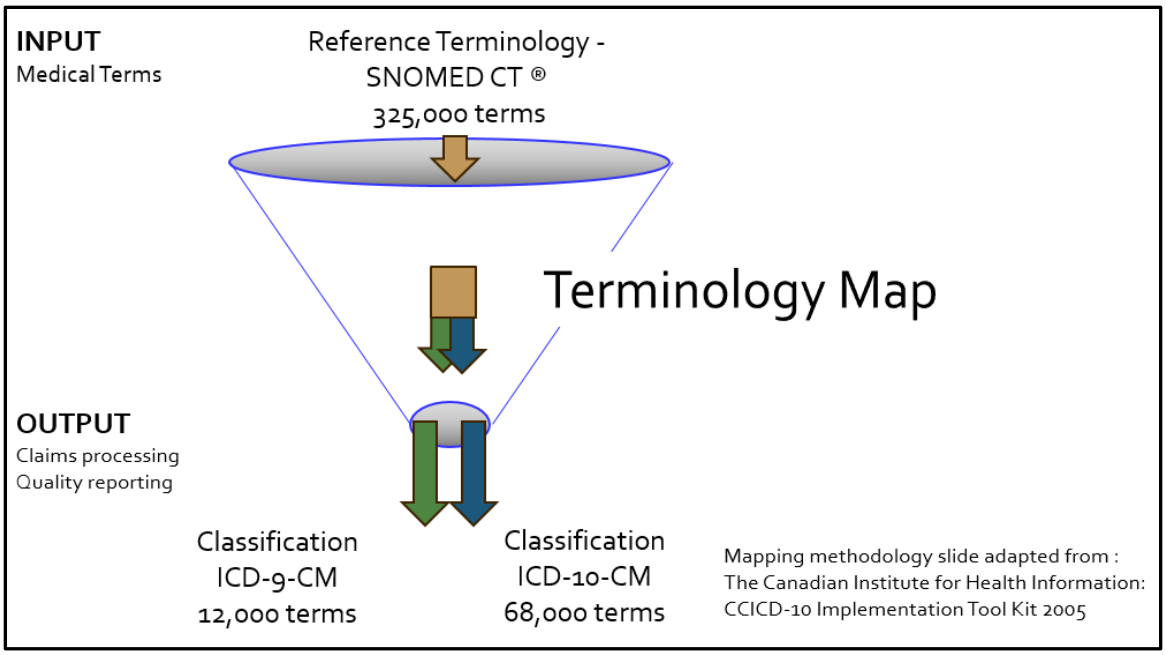


Figure 4. Conceptual diagram of terminology mapping (adapted from Bowman, S. E., 2005).

A conceptual diagram of the application of diagnosis codes to patient records is in Figure 5. The clinician selects one of many clinical terms supplied by the third party terminology vendor. The local diagnosis code associated with the selected term is applied to the patient problem list or encounter diagnosis. The code mappings link the local codes to the standard terminologies, ICD and SNOMED.

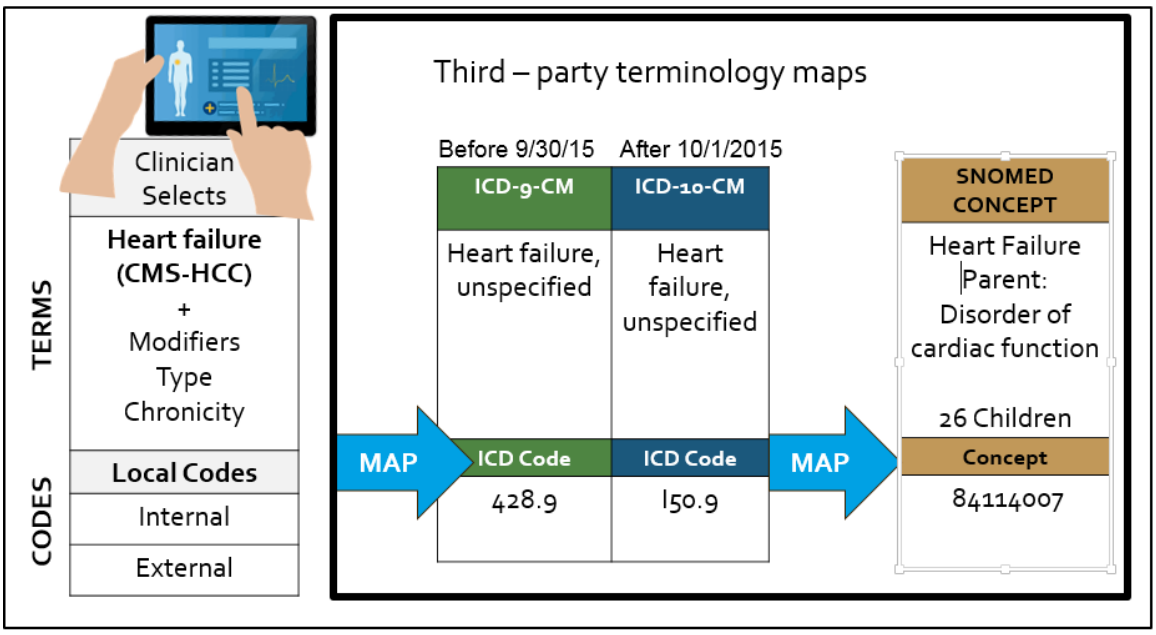


Figure 5. Conceptual diagram of the application of diagnosis codes to patient records

Intelligent Medical Objects, Inc. (IMO®) offers a terminology solution that maps medical terms commonly used by clinicians to ICD and SNOMED. Epic, Cerner, NextGen, and several other EHR systems incorporate IMO terms in their software. (Kottke & Baechler, 2013) Apelon Distributed Terminology System (DTS) is an open source solution. As a licensee of IHTSDO, the U.S. National Library of Medicine (NLM) is the single public source of SNOMED CT data in the United States (U.S. National Library of Medicine, 2011). The internal mapping relationships between local and standard diagnosis terminologies support the output, retrieval and aggregation of patient cohorts using ICD-9-CM, ICD-10-CM or SNOMED CT (Figure 6). Terminology middle-ware functions as an interface terminology employing internal mappings to shield users from the need to assign standard terminologies such as ICD and SNOMED codes directly. A survey of SNOMED users found, “In most cases SNOMED CT had been so seamlessly integrated

that that users were unaware that they were using SNOMED CT through and interface terminology," (Lee et al., 2013).

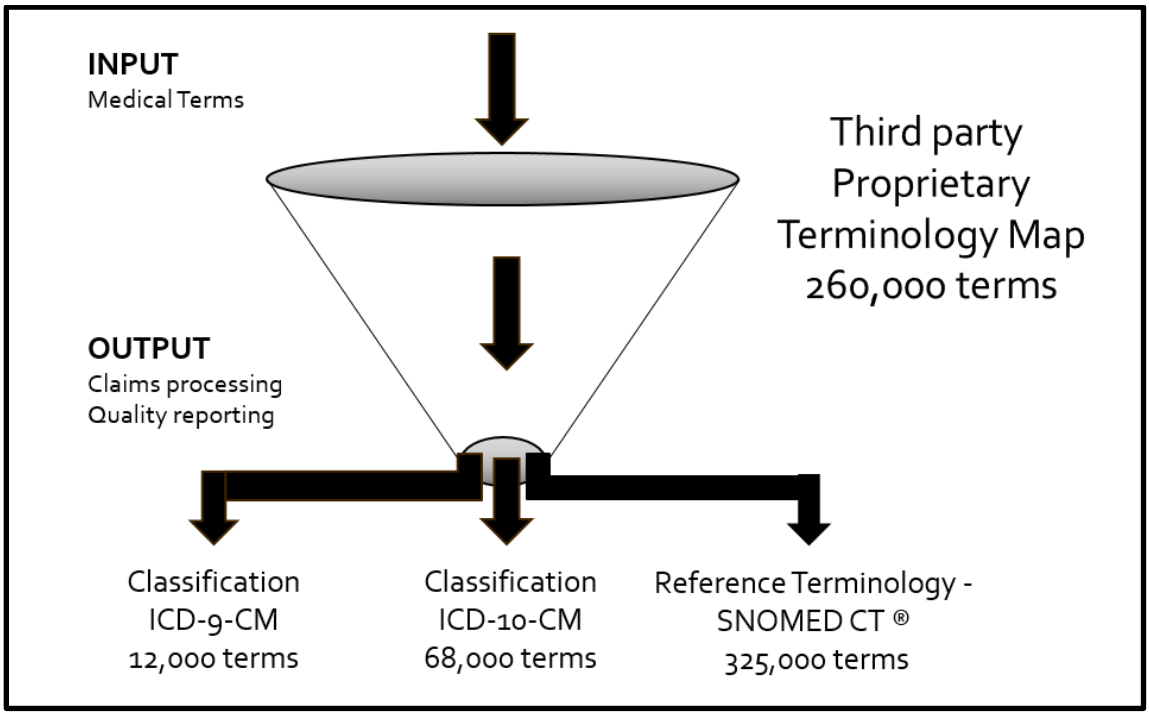


Figure 6. Conceptual diagram of terminology mapping circa 2017

Challenges encountered when mapping International Classification for Nursing Practice (ICNP) to SNOMED-CT included inconsistencies, redundancies, and deficiencies of SNOMED CT concepts (Kim, 2016). The consequence of attendant errors are acknowledged in the American Health Information Management Association publication, Data Mapping and its Impact on Data Integrity.

> …poorly designed or out-of-date mappings create significant data integrity problems in health information systems. Undetected errors in data maps have the potential to introduce many problems including the filing of false claims to insurers, delivering the wrong information for patient care and/or quality

measures, or causing a breach in patient privacy (Hyde, Rihanek, Santana-Johnson, & al, 2013).

The most serious problems may actually derive from delivery of the wrong care to a patient or the failure to deliver intended interventions.

1.5 Maintenance of Terminologies and Maps

Terminology maintenance is a necessary and expected activity in the domain of health information technology (HIT) due to rapid evolution of the evidence base, regulation, clinical and administrative practices (Cimino, Clayton, Hripcsak, & Johnson, 1994). Saitwal et al. (2012) state "*… all mappings must be maintained and updated as errors are found and corrected, and as the source and target terminologies change."* Terminology updates and correlated cross-terminology mapping revisions occur in all modern EHR systems, and yet there is no standardized way to identify and measure the impact of this activity on downstream uses of the diagnosis codes. Even when a cross-terminology map has demonstrated good performance, it must be maintained to stay current if concepts and terms are added, removed, or the meaning of a code changes, so as to continue to produce consistent results (Rea et al., 2012). Codes may also be deprecated, meaning they are no longer active to be newly applied to patient records, but remain in the EHR for backward compatibility. An important finding by the Strategic Health IT Advanced Research Projects Area 4 Consortium is that, "The SHARPn demonstration did not deal with multiple versions of terminologies or updates to terminologies, but it became apparent that any robust data normalization effort will need to do so," (Rea et al., 2012).

When the U.S. transitioned from ICD-9 to ICD-10 for cause of death reporting, the CDC provided preliminary comparability ratios to indicate the extent of discontinuities resulting from the coding changes (Anderson, Miniño, Hoyert, & Rosenberg, 2001). Bridge-coding studies are designed to measure the effects of new terminology revisions using duel coded datasets. According to Fenton and Benigni (2014), "Longitudinal data comparisons can only be reliable if they use comparability ratios or factors which have been calculated using records coded in both classification systems." However, it's important to recognize that the comparability statistic only describes the net change in the resulting cohorts, and thus lacks detail as to which patients were added and deleted as a result of the terminology changes. Figure 7 illustrates various methods for evaluating shifts caused by terminology mapping. Comparison of the overlap of terms only as in diagram A will provide an indication of the terms added and dropped from a phenotype definition due to terminology mapping, but the effect of those changes on the patient population depends on the prevalence of the codes, and can therefore not be established by this method. Diagram B shows the method of instantiation whereby the mapping difference may be stated by comparing the cohorts of disease positive patients found by each diagnosis terminology. A drawback of this method is that although the difference can be ascertained, there is no way to determine if the change is better or worse. The method in diagram C uses a reference standard as a common comparator from which to find true positives and establish a rate of error. Comparison of the phenotype performance to the reference standard provides a quantitative method by which to establish which terminology performs better.

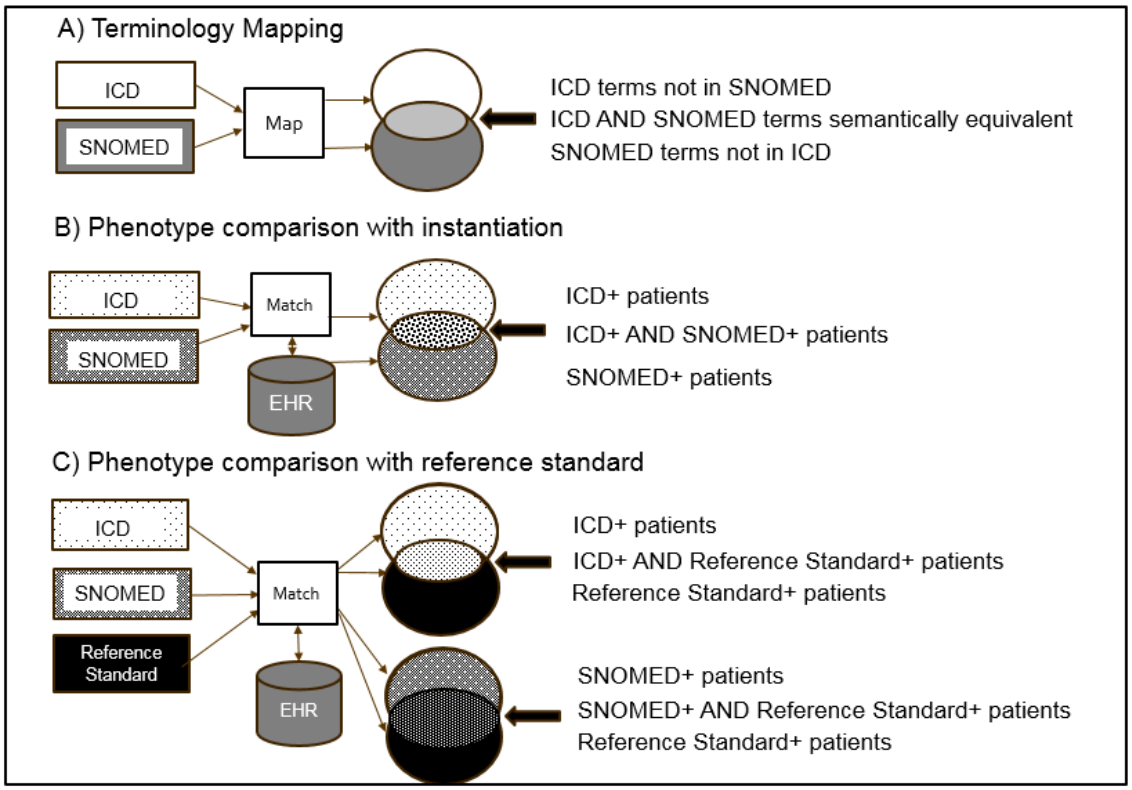


Figure 7. Evaluating the difference between diagnosis terminologies

Fluctuations of membership in a population health CKD registry have been observed following terminology updates to the EHR system. The January 28-31, 2016 EHR terminology map maintenance at a university health system that applied SNOMED updates to the EHR system resulted in the decrease of the CKD registry population by 2,065 patients, roughly one third of its members. This unexpected change was caused when the EHR internal mapping from *ICD-9-CM 585 Chronic Kidney Disease (CKD)* to *SNOMED 433146000 Chronic kidney disease stage 5* was inactivated. A data integrity process monitoring the number of patients in this registry identified the unintended consequence of this update, and the problem was resolved by changing the inclusion rule to ICD-10-CM coding exclusively.

Due to the size and complexity of the current clinical terminologies, maps are likely to have some errors. Terminology maintenance creates a dynamic setting in which the adjustment of a map or the failure to do so may introduce errors into the phenotype algorithms that rely on mappings. This effect was reported when ICD-9-CM diagnosis codes for acute liver failure were mapped from ICD to SNOMED CT (Reich et al., 2012). The acute renal failure cohort was less than 10 patients, and the resulting shift of 1 patient was too small to detect the true rate of change.

1.6 Registry Membership is a Phenotyping Task

In order to measure the properties of a phenotype algorithm, the true state of each patient must be established using a reference standard. The generalized 2 X 2 contingency table with experimental results True positive, False Positive, True Negative, and False Negative can be used to calculate a number of additional summary statistics (Table 3). Sensitivity, specificity, and positive predictive values (PPV) are commonly reported measures of phenotype performance. When selecting the best version of a phenotype algorithm, there is typically a trade-off between finding all of the patients expected to benefit from an intervention and excluding patients who won't benefit or may be harmed by an intervention. The false discovery rate (FDR) and the false omission rate (FOR) are of particular concern because these statistics predict the number of patients inappropriately included or exclude from intervention group in error. The review of the literature provided deeper insight into the use of these statistics in the phenotyping literature.

Table 3. *Summary statistics derived from a 2 X 2 contingency table*

		Reference Standard		Total
		Disease	No Disease	
Phenotype Algorithm	Positive	True Positive (TP)	False Positive (FP)	Algorithm Positive (AP)
	Negative	False Negative (FN)	True Negative (TN)	Algorithm Negative (AN)
Prevalence=DP/N	Accuracy = (TP+TN)/N	Disease Positive (DP)	Disease Negative (DN)	Population = N
PPV=TP/AP	NPV=TN/AN	True Positive Rate, Sensitivity, Recall TPR =TP/DP	False Positive Rate FPR =FP/DN	
False Omission Rate FOR=FN/AN	False Discovery Rate FDR=FP/AP	False Negative Rate FNR=FN/DP	True Negative Rate, Specificity TNR=TN/DN	

Chapter 2: Literature Review

A PubMed literature review was conducted May 11-22, 2016 with the help of a research librarian at the University of California, San Diego (Mary Wickline). The search strategy was designed to target articles about electronic phenotype studies (ICD or SNOMED) for diabetes, CKD or heart disease (Figure 8). The final combination of search terms yielded 222 abstracts that were evaluated on the following inclusion criteria:

- data source is observational clinical data and coding terminology is ICD-9-CM, ICD-10-CM or SNOMED-CT
- diagnosis phenotype accuracy is reported
- use case was chronic disease management or quality measures for diabetes, CKD, or heart disease
- article in English and full text available

The full text articles meeting the criteria (91) were coded for references to ICD and SNOMED terminologies (54 unique articles), electronic phenotyping methods (61 unique articles) and disease classification (44 unique articles). Appendix A contains the details of the search terms.

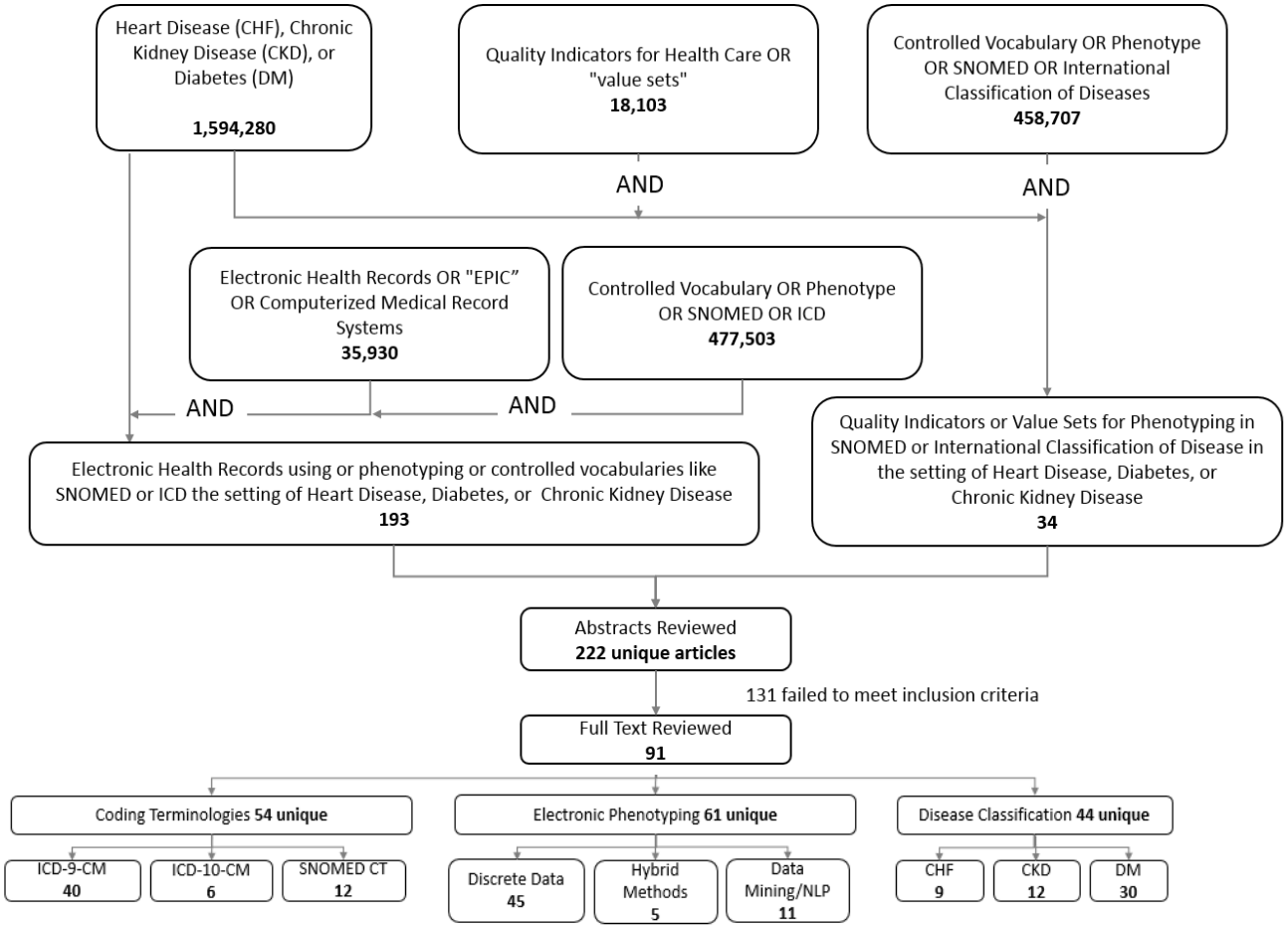


Figure 8. Literature search strategy

2.1 Current Uses of ICD and SNOMED Medical Terminologies

The articles about ICD and SNOMED were reviewed to glean information on the frequency and context of the diagnosis coding options. Of these 54 articles, 49 (74%) used ICD-9-CM, 12 (22%) used SNOMED CT and 6 (11%) used ICD-10-CM. Several articles compared terminologies, and included more than one.

ICD-9-CM was primarily developed for billing and administrative applications and does not necessarily imply a well-defined robust and logical hierarchy for the codes. (Pathak, Kiefer, Bielinski, & Chute, 2012) However, clinical researchers often use ICD billing diagnosis codes for phenotyping because these codes are mandated for payment within the U.S. healthcare system, and the disease, signs and symptoms in ICD terms are often used as a surrogate for the disease phenotype (Pathak et al., 2013); (Schildcrout et al., 2010). ICD-9 was used as a primary phenotype rule in studies diabetes, (Fort, Wilcox, & Weng, 2014) (Klompas et al., 2013) (Meyers, Candrilli, & Kovacs, 2011) (Nag et al., 2007) (Wilke et al., 2007) (Zhong et al., 2016) kidney disease (Ferris et al., 2009[[Brieler, 2016 #1397) (Navaneethan et al., 2011) (Murff et al., 2011) (Cipparone et al., 2015) and heart disease (Baker et al., 2007) (Broberg et al., 2015) (Floyd, Blondon, Moore, Boyko, & Smith, 2016) (Kleinberg & Elhadad, 2013) (Hoang et al., 2014) (Udris et al., 2001) These articles, dated 2001-2016 use relatively simple ICD-9 algorithms, using the character X to denote a place holder for any digit, for example "Diabetes 250.X0, 250.X2" (Brieler, Lustman, Scherrer, Salas, & Schneider, 2016; Broberg et al., 2015); (Andrade et al., 2011).

Recent adoption of ICD-10-CM in October 2015 helps explain the low number of articles retrieved in May of 2016 as compared to ICD-9-CM and SNOMED. The mandated coding standard had yet to be fully explored in the literature at the time of the search. Canada adopted ICD-10 in 2001, well before the United States. So, Evans, and Quan (2006) compared the performance of ICD-9 and ICD-10 in the retrieval of nine AMI comorbidities, and found similar sensitivity, specificity, positive predictive value and negative predictive value. This study included 193 patients with known AMI as confirmed by chart review, and also considered prevalence and mortality in a Canadian province from 1994-2004. So et al. (2006) concluded that ICD-10 coding algorithms performed similarly to ICD-9. Although ICD-10 was first published over a decade ago, the ICD-10-CM U.S. extension has over 3 times the codes of its ICD-10 parent terminology. It will take time to collect longitudinal ICD-10-CM data, and to develop and validate new ICD-10-CM coding algorithms.

SNOMED CT has an ontological structure that was more frequently correlated with natural language processing phenotyping. de Keizer et al. (2000) define an ontology as "a specification of concepts, relations and functions for a domain". The relations convey lexical relationships (shared meaning) between terms that are semantically different, for example "kidney disease" and "renal failure". Liaw et al. (2014) conclude that integrating multiple data elements with an EHR using ontology-based case-finding algorithms can improve the accuracy of a Type 2 Diabetes Mellitus registry. SNOMED CT-AU was the domain ontology used in this Australian study. Although it's true that data fragmentation and inaccuracy can negatively impact the quality of phenotype results, (Wei, Leibson, Ransom, Kho, & Chute, 2013) (Jolly et al., 2014) the Liaw study does not separate the

effects the ontology versus the effects of the inclusion of additional discrete items in a discrete rule-based approach.

A case has been made that SNOMED© CT simplifies querying of clinical data to the extent that knowledge of clinical medicine, coding schemes and database structure is no longer required (Lieberman, Ricciardi, Masarie, & Spackman, 2003). It may seem so when myocardial infarction, coronary artery disease, heart failure and hypertension can each be mapped to a single SNOMED concept. This study was conducted without the use of a reference standard, and ICD-9 codes were assumed to designate the "true" state. The reported Recall rate for Type II Diabetes Mellitus was 0.987 and Heart Failure was 0.921. However, lack of detail about the methods of SNOMED mapping and failure to compare against a reference standard lead to questions about the reproducibility of the results. A concerning finding was that the concept of 'insulin dependent diabetes mellitus' was not included under the type I diabetes hierarchy in SNOMED which explained the recall rate of 0.741 for Type I Diabetes Mellitus (Lieberman et al., 2003). It is precisely these types of mapping decisions that must be studied across a wide range of disease states.

Ultimately, the use of SNOMED CT in CMS quality programs will likely be the most significant driver of adoption. The Meaningful Use Stage 2 rule identifies SNOMED CT as a clinical terminology standard of certified EHR systems. Therefore, it is critical to understand the principles and implications of using SNOMED CT and other clinical standards for knowledge representation within EHR systems (Monsen et al., 2014).

2.2 Electronic Phenotyping Methods

The review of 61 articles on phenotyping methods revealed that a nearly three quarters (74%) used discrete data, as compared to free text data (18%) or hybrid techniques (8%).

Some argue that a gold standard is required to evaluate the retrieval performance of a terminology (Brown & Sonksen, 2000). Rubbo, Fitzpatrick, Denaxas and colleagues contend that a major problem in evaluating studies of EHR-derived diagnoses is the implementation of a "gold standard" (Rubbo et al., 2015). The method is time intensive, and depending on the use case, some suggest that any baseline standard will suffice as a comparator (Agarwal et al., 2016). Agarwal, Podchiyska, Banda and colleagues are among the growing number of researchers who are developing improved automation for the creation of gold standards from clinical sources, and they have found value in the use of a semi-automated “silver standard” for labeling training sets for phenotype models. Stanfill et al. (2010) reported, literature evaluating automated coding and classification systems only reports this step in approximately 50% of studies comparing performance of a coding terminology to a gold standard.

To achieve the high levels of accuracy required to drive prospective chronic disease care, population health registries must look beyond the readily available discrete diagnosis codes, and include diagnosis, lab values, procedure results, and so on (Wei et al., 2016). Since population health interventions trigger actionable intervention on real patients, the registry inclusion rule (functioning as a phenotype algorithm) routinely filters out deceased patients and those with no medical visits or acute care in the last 3 years. Active patients are grouped according to diagnoses for chronic disease management. Navaneethan and colleagues from Cleveland Clinic implemented an EHR-based CKD registry using an inclusion rule based on one face-to-face encounter, two encounter diagnoses for CKD, and/or two estimated (eGFR) values indicating CKD stage 5 or higher (Navaneethan et al., 2011). The CKD inclusion rule was approved following a

chart review of 20 randomly selected charts by three reviewers, and the resulting registry of over 57,000 patients has become a valuable research for tool for studying CKD comorbidities (Navaneethan et al., 2011).

Studies using discrete data most frequently reported sensitivity, specificity, and positive predictive value (PPV). Data mining techniques reported precision and recall. Despite the different naming conventions precision and recall formulas are the same as PPV and sensitivity. These two statistics were reported by more than 50% of the articles and specificity by about 30%. Other statistics were reported less than 20% of the time including; negative predictive value, 2 x 2 contingency table (or text equivalent), accuracy, area under the curve (AUC), receiver operating curve (ROC), Chi square, and percent match (Figure 9). Overall, there was little consistency in the published evaluation methods or performance measurements for phenotypes. Appendix B Electronic Phenotyping Evaluaton Methods lists the reviewed articles referencing each statistic.

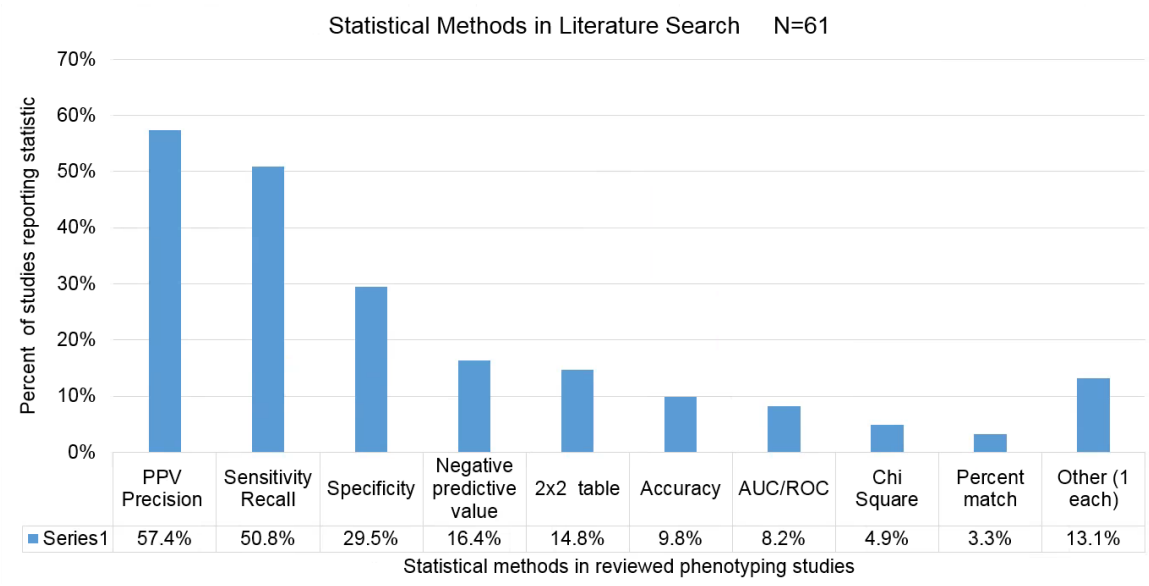


Figure 9. Statistical methods used in 61 phenotyping articles from literature search

Reich et al. (2012) explored the impact of terminology changes on queries of the Observational Medical Outcomes Partnership research network data. Reich's results clearly show the number of patient records lost or gained as compared to the original ICD-9-CM cohort when mapped to SNOMED and MedDRA definitions. Of course, inclusion of the 2x2 contingency table in published results would support the calculation of all the summary statistics listed in Table 5 including Sensitivity, Specificity, and PPV, but only 15% of the articles provided this information. True Positive, False Positive, True Negative, and False Negative values provide valuable information even when presented in non-standard formats. Garvin et al. (2013) reported these values in a single table combined with sensitivity, specificity and PPV.

2.3 Disease classification for Diabetes, Chronic Kidney Disease and Heart Failure

My review of the literature found additional information about the significance and interrelatedness of diabetes, CKD and heart failure. Of the 44 phenotype articles mentioning a disease condition of interest, diabetes was by far the most heavily reported (30, 68%) followed by CKD (12, 27%) and heart failure (9, 20%).

Diabetes has been rapidly increasing in prevalence in recent years. CDC estimates 1.7 million new adult cases of Type 2 diabetes are diagnosed each year (Centers for Disease Control and Prevention, 2015). If the trend continues 438 million adults are estimated to develop diabetes by the year 2030 (Pathak et al., 2012; Rathmann W & Giani G, 2004). Type 2 diabetes accounts for 90-95% of all new diabetes cases in the U.S. People with diabetes typically experience healthcare costs 2.3 times higher than non-diabetics. Approximately 40% of CKD cases are attributable to diabetes (Huopaniemi et al., 2014; Meyers et al., 2011; Nadkarni et al., 2014). The prevalence of CKD in the U.S. is

estimated to be 13% (Navaneethan et al., 2011). CKD is a precursor to End Stage Renal Disease (ESRD) that warrants dialysis or transplantation (Navaneethan et al., 2013; Schroeder et al., 2015). Even small degrees of renal impairment are associated with increased cardiovascular disease risk, cardiovascular mortality, and health care costs. The status of kidney function is based upon a calculation that includes a laboratory value from routine metabolic profile called serum creatinine. This value is used in the calculation of eGFR which is used to assess the severity of the condition (Levey & Coresh, 2012). Table 4 shows the eGFR ranges for Chronic Kidney Disease Stages 1 through 5. Stage 5 is the most severe kidney disease which can only be treated by dialysis or transplantation, and is also known as End Stage Renal Disease (ESRD). Note that ESRD is often equated with and typically commences in Stage 5, but only refers to patients starting or receiving dialysis or transplantation (United States Renal Data System, 2017).

Table 4. *Chronic Kidney Disease staging 2014 USRDS Annual Data Report*

Stage 1	eGFR $\geq$ 90 mL/min per 1.73 m^2
Stage 2	eGFR 60–89 mL/min per 1.73 m^2
Stage 3	eGFR 30–59 mL/min per 1.73 m^2
Stage 4	eGFR 15–29 mL/min per 1.73 m^2
Stage 5	eGFR < 15 mL/min per 1.73 m^2

Heart failure causes shortness of breath, weight gain and tiredness when the heart is unable to supply sufficient blood flow to the body. The primary diagnostic test is cardiac

ejection fraction also known as Left Ventricular Ejection Fraction (Bielinski et al., 2014). A LVEF measure less than 40 is the established clinical definition of heart failure which can be either chronic or acute. Agarwal et al. (2016) warns that guidelines and quality measures for heart failure will need to account for multiple measures of LVEF that may change over time, with a patient moving across the threshold for heart failure in both directions. About half of the people who develop heart failure die within 5 years of diagnosis. (Mozaffarian D et al.)

Chapter 3: Aim 1. Quantify impact of terminology choice (ICD vs SNOMED)

3.1 Introduction

The statistical analysis of sensitivity and specificity is dependent on the population, and cannot be assumed to be consistent across multiple EHR instances or disease states. It is therefore fundamental to establish a methodology by which the retrieval properties of a phenotype rule are established locally to inform clinical and financial decision making. Terminology maps supplied by third-party middleware providers enable rule-based retrieval of patient diagnoses coded in ICD and SNOMED terminology. The Value Set Authority Center (VSAC) provided matched sets of chronic disease value sets coded in ICD and SNOMED which were used to isolate the effect of terminology choice in the retrieval of diabetes, CKD and heart disease. As a licensee of IHTSDO, the U.S. National Library of Medicine (NLM) is the single public source of SNOMED CT data in the United States (U.S. National Library of Medicine, 2011). VSAC is the repository for official versions of diagnosis value sets for regulatory quality programs such as Meaningful Use and Clinical Quality Measures. The value sets are maintained by the National Library of Medicine (NLM), in collaboration with the Office of the National Coordinator for Health Information Technology (ONC) and CMS (Bodenreider et al., 2013). When the VSAC value sets are used in a phenotype algorithm, the selection of a patient cohort is like a binomial diagnostic test, rendering a positive or negative result for each disease state. When both ICD and SNOMED phenotype algorithms "test" the same patient population, the study design is paired, and McNamar's test for dependent

proportions is recommended for these conditions (Zhou, Obuchowski, & McClish, 2011b).

3.2 Methodology

The null hypotheses for Aim 1, $H_{0DM,}$ states that for diabetes, CKD and heart disease, the sensitivity and specificity of phenotype algorithms using ICD-10-CM and SNOMED will be the same (Equation 1A). The alternative hypothesis, $H_{ADM,}$ states that for each disease, the sensitivity and specificity of phenotype algorithms using ICD-10-CM and SNOMED exclusively will be different (Equation 1B).

Equation 1. Aim 1 hypotheses for comparison of ICD and SNOMED value sets

A. Null hypothesis

$$H_{0DM}: \theta ij\ Phenotype(ICD) = \theta ij Phenotype(SNOMED)$$

B. Alternative hypothesis

$$H_{ADM}: \theta ij\ Phenotype(ICD) \neq \theta ij Phenotype(SNOMED)$$

Where θij is the true summary measure of the *i*th Disease D_i, *i=1-3 and jth* Summary Measure M_j, *j*=1-2 for each phenotype algorithm.

Disease D= {diabetes, CKD, heart failure}

Summary Measure M= {Sensitivity, Specificity}

Study data included retrospective observational data collected in an Epic® EHR during care delivery. The study included alive patients age 18 or older as of September 30, 2013 with at least one arrived or completed office visit encounter between October 1, 2013 and September 30, 2016. The office visit criterion was to limit the study to patients who were

seen in the ambulatory setting, and are therefore candidates for chronic care management. Study datasets for each disease state were extracted from the Epic® Clarity reporting database, and contained demographics, diagnoses, laboratory results, and procedure results to inform the reference standard phenotypes for diabetes, chronic kidney disease and heart failure. The population represents the actual distribution of ethnic and racial backgrounds, and gender served by UCSDH as no exclusion was made on the basis of gender, race or ethnicity, pregnancy status, or sexual orientation. The study population contained women of child-bearing potential, but the pregnancy status for individual patients was not ascertained. Prisoners who received care at UCSDH were included, but the investigators had no way of identifying which subjects were prisoners. Cancer patients were also included, but cancer diagnoses codes were not relevant to the study, and were not captured in the study data.

3.2.1 Phenotypes. Three types of phenotype rules for each disease were implemented as SQL queries to extract patient cohorts from the research data: ICD diagnosis only; SNOMED diagnosis only; and a reference standard based on research phenotypes (Figure 10).

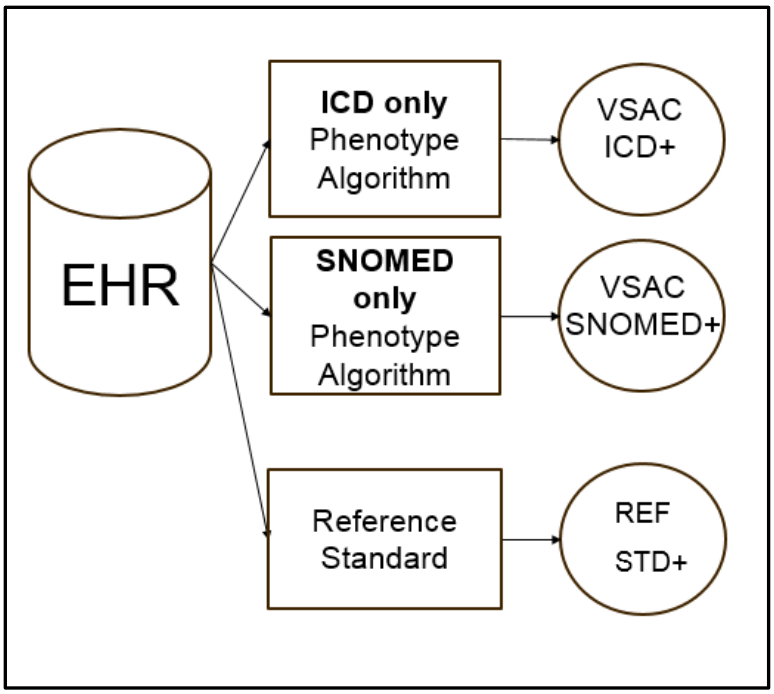


Figure 10. Aim 1 Patient cohorts for ICD, SNOMED, reference standard phenotypes

VSAC value sets for VSAC Value Sets for Diabetes, Chronic Kidney Disease, Stage 5, and Heart Failure were retrieved on May 22, 2016. See Appendix C Value Set Authority Center Downloads for details. The Grouper Editor is the EHR activity for creating and editing value sets within EpicCare Ambulatory 2015®. An embedded utility in the Grouper editor was used to resolve the standard ICD and SNOMED terminology codes into value sets of local diagnosis codes. Each local code has its own mapping to ICD-9-CM, ICD-10-CM, and SNOMED CT. Hence, the internal terminology map was externalized. Patient cohorts were identified when the local diagnosis ID matched a local diagnosis in a patient's problem list or office visit encounter history. The VSAC diabetes value set contained 146 ICD-10-CM codes that mapped to 9,328 local codes. The utility converted 36 SNOMED codes for diabetes to 7,118 local codes. Diabetes has the largest

number of local codes of the diseases studies, but it not uncommon for value sets to include thousands of local codes.

Table 5 provides examples of some of the complicating conditions encountered with the code mappings. 1) *Diabetes mellitus complicating pregnancy* is a temporary condition generally excluded from chronic care management. 2) *Diabetes mellitus with HbA1C goal between 7 and 8* blurs the distinction between controlled and uncontrolled by setting an explicit goal. 3) *Type 1 diabetes mellitus with peripheral angiopathy without gangrene* is mapped to two SNOMED CODES, 127014009 and 46635009. 4) Local code for *Non-insulin dependent diabetes mellitus* has no ICD-10-CM mapping. 5) *Insulin dependent type 2 diabetes mellitus, controlled* is an example of why diabetic laboratory results can't be used to determine the type of diabetes. The term, "controlled" implies that the diabetic patient would have normal readings on glucose and HgbA1c tests. 6,7,8) Depending on the phenotype use case, granular descriptions of diabetes complications may or may not be meaningful. If a clinician judges a term to be inappropriate for the intended use of the data, it may be difficult to remove an individual term. For instance dropping ICD-10-CM code E11.9 will remove 2) *Diabetes mellitus with HbA1C goal between 7 and 8* and 5) *Insulin dependent type 2 diabetes mellitus, controlled.*

Note that clinicians may not have visibility of all of these local code choices. Some are marked as clinically inactive and some have never been applied to patient records. However phenotyping algorithms routinely test whether any of the resolved local codes appear on patient records.

Table 5. *Local diabetes code mapping examples*

Local Code	Term	ICD-9	ICD-10	SNOMED CT
Example 1	Diabetes mellitus complicating pregnancy	648.00 250.00	O24.919	609496007
2	Diabetes mellitus with HbA1C goal between 7 and 8	250.00	E11.9	365845005
3	Type 1 diabetes mellitus with peripheral angiopathy without gangrene	250.71 443.81	E10.51	127014009
4	Non-insulin dependent diabetes mellitus	250.00		44054006
5	Insulin dependent type 2 diabetes mellitus, controlled	250.00 V58.67	E11.9 Z79.4	237599002
6	Type 2 diabetes mellitus with left diabetic foot ulcer	250.80 707.15	E11.621 L97.529	1521000119100
7	Type 2 diabetes mellitus with diabetic cataract	366.41		420756003
8	Type 2 DM w establish diabetic nephropathy	250.40		420279001

Public research phenotypes with established high sensitivity and specificity were adapted for use as a reference standard for diabetes and chronic kidney disease. These algorithms used discrete quantitative values in the medical record to define a disease state, such as diagnoses, medications, laboratory values, and/or procedure findings. Although the development of a reference standard was a necessary step in this study methodology,

defining the best methods for reference standard development is outside the scope of this research.

The Type 2 Diabetes Mellitus phenotype developed by the eMERGE network was used as the reference standard with limited modifications (Pacheco, 2012).

This algorithm had 98.2%-100% Positive Predictive Value (PPV) when applied across institutions (Northwestern University, Vanderbilt University, Marshfield Clinic) (Kho et al., 2012). Based on the reported data from Kho's table 3, I calculated sensitivity (.995) and specificity (.986) of the algorithm over the 3 sites combined (N=350).

A number of challenges were discovered in adapting the research algorithm for use in the population health context which are reported in Appendix D: Reference Standard for Diabetes.

Table 6. *eMERGE Diabetes Type 2 Case Inclusion Rules*

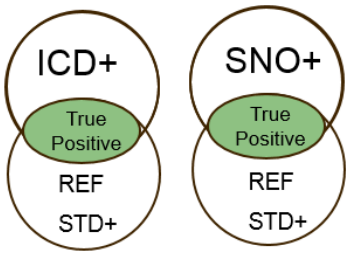
eMerge Phenotype Rule	Type 1 diabetes DX	Type 2 diabetes DX	Type 1 diabetes Med	Type 2 diabetes Med	Type 2 Med prescribed before Type 1 Med	Abnormal diabetes Labs
1	No	Yes	Yes	Yes	Yes	
2	No	Yes		Yes		
3	No	Yes				Yes
4	No			Yes		Yes
5	No	Yes	Yes			

Note. Patient was included in diabetes reference standard cohort if any rule was met

An eMERGE network phenotype was also selected for the CKD population. Nadkarni et al. (2014) have reported a phenotype algorithm for CKD with a Positive Predictive Value of 95.95 (CI 90.85-95.08) and a Negative Predictive Value of 93.25 (CI 90.85-95.08). CKD patients were selected based on a diagnosis of CKD, kidney transplant, or other kidney disease including renal failure and dialysis. Alternatively two laboratory measures of GFR less than or equal to 60 over a period of 90 days or more confirmed a CKD diagnosis. All CKD patients identified by diagnosis or lab, were then subject to a final test of eGFR less than 15 to limit selection to CKD Stage 5.

The reference standard for heart failure identifies patients with ICD-9-CM or ICD-10-CM code for heart failure in one active problem list diagnosis or two encounter diagnoses or evidence of LVEF less than or equal to 40%.

3.2.2 Statistical Analysis. In order to fully understand how terminology mapping affects cohorts, the coded diagnosis terms must be instantiated, i.e., programmatically matched to EHR data. The unique study IDs of the ICD cohort were compared with patient IDs in the reference standard, likewise the SNOMED cohort was compared to the reference standard. The resulting True Positive, False Positive, True Negative and False Negative values were recorded in 2x2 contingency tables as shown conceptually in Figure 11. These values were used to derive the sensitivity and specificity of the ICD and SNOMED diagnosis phenotypes for each disease.



		Reference Standard		Total
		Disease	No Disease	
ICD	+	**True Positive (TP)**	False Positive (FP)	ICD +
	-	False Negative (FN)	True Negative (TN)	ICD -
SNOMED	+	**True Positive (TP)**	False Positive (FP)	SNOMED +
	-	False Negative (FN)	True Negative (TN)	SNOMED -

Figure 11. Conceptual diagram of the transformation of ICD, SNOMED and Reference Standard Cohorts into 2x2 contingency tables.

The null hypothesis states that the sensitivity and specificity of the chronic disease phenotype algorithms would be the same whether an ICD or SNOMED value set is used. The McNamar's test statistic, χ^2 for sensitivity was calculated from the set of patients with reference standard positive. Using Zhou et al. (2011b) notation, m_{111} is the number of patients with both tests positive and m_{101} is the number with the first test negative and the second test positive. The χ^2 statistic for specificity is calculated from the set of patients with reference standard negative. These values must be calculated directly from the test data, and cannot be derived from a 2x2 contingency table. McNamar's test for dependent proportions was performed using SPSS version 24 (IBM, 2016) to evaluate the statistical significance between the sensitivity and specificity of ICD and SNOMED diagnosis phenotypes.

Equation 2. McNamar's χ^2 statistics used to establish the significance of differences in

sensitivity and specificity between ICD and SNOMED phenotype algorithms.

Sensitivity: for patients with **positive** Reference Standard

$$\chi 2 = \frac{(m_{110} - m_{101})^2}{m_{110} + m_{101}}$$

Specificity: for patients with **negative** Reference Standard

$$\chi 2 = \frac{(m_{010} - m_{001})^2}{m_{010} + m_{001}}$$

3.3 Findings

The study population included 201,917 patients age 18 or older with at least one office visit during the study period. The demographics in Table 7 Age and sex of study population represent the actual EHR population at that time. Percent of patients in each age bracket was evenly distributed from 18-49 (18-29, 16.9%; 30-39, 16.3%; 40-49, 15.6%). The age distribution peaked between age 50-70 (50-59, 19.7%; 60-69, 17.0%), the rapidly dropped off over age 80 (80-89, 4.1%; 90+, 0.7%). Gender was skewed toward female (57.5 female vs 42.5% male). The missing information on sex was not significant.

Table 7. *Age and sex of study population*

Age	Frequency	Percent
18-29	34,112	16.9
30-39	32,918	16.3
40-49	31,598	15.6
50-59	39,759	19.7
60-69	34,397	17.0
70-79	19,567	9.7
80-89	8,234	4.1
90+	1,332	0.7

Sex	Frequency	Percent
Female	116,035	57.5
Male	85,880	42.5
Missing	2	
Total N=	201,917	

Table 8 Race and ethnicity of study population was predominately white (61.0%). A majority (76.5%) reported their ethnicity as "not Hispanic or Latino." Legacy ethnicity data of "African American" (306), "American Indian/Eskimo (18) "Asian/Pacific Islander" (414), and "Caucasian" (3,014) were included in the count of "not Hispanic or Latino". The count of Unknown included Multi-Racial (971). There were 81 missing values which were not significant.

Table 8. *Race and ethnicity of study population*

Race	Frequency	Percent
American Indian or Alaska Native	905	.4
Asian	19,398	9.6
Black or African American	8,558	4.2
Native Hawaiian or Other Pacific Islander	828	.4
Missing	1,428	.7
Other Race or Mixed Race	36,305	18.0
Unknown (Patient cannot or refuses to declare race)	11,348	5.6
White	123,147	61.0

Ethnicity	Frequency	Percent
Not Hispanic or Latino	154,369	76.5
Hispanic or Latino	32,860	16.3
Unknown	14,607	7.2
Missing	81	0.0
Total	201,917	

The details of the study population are for informational purposes only. Specific clinical findings cannot be generalized to other populations, but the techniques for quantifying phenotype performance are generalizable to the extent that they can be applied to any population.

For diabetes, Table 9 shows the VSAC ICD and SNOMED value set phenotype performance. ICD-10-CM outperformed SNOMED CT with higher true positives (9,345, 7471), lower false negative (1,934, 3808). ICD-10-CM was worse than SNOMED CT with lower true negatives (185,285, 185,926), and higher false positives (5,349, 4,708).

Overall, ICD-10-CM had greater sensitivity, and worse specificity. For diabetes, McNamar's $\chi 2$ (1,541.357) is greater than the critical value for sensitivity, and the null hypothesis is rejected. McNamar's $\chi 2$ (149.653) is also greater than the critical values, and the null hypothesis is rejected for specificity. For the VSAC diabetes value sets, the difference in sensitivity and specificity is significantly different for ICD-10 versus SNOMED (p=0.001). ICD-10-CM sensitivity was better than SNOMED by 16.7%, and the sensitivity was worse than SNOMED by 0.3%.

Table 9. *Diabetes: VSAC ICD and SNOMED value set phenotype performance*

Diabetes Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
ICD-10-CM	Yes	9,345	5,349	14,694	0.829	0.972
	No	1,934	185,285	187,219		
SNOMED-CT	Yes	7,471	4,708	12,179	0.662	0.975
	No	3,808	185,926	189,734		
Total		11,279	190,634	201,913		
(McNamar's test statistic $\chi 2$)					**1,541.357**	**149.653**

Critical value 10.827
df=1, level .001

Note that ICD-10-CM overall false positive and negative errors (7,283) were fewer than SNOMED (8,516), and that the number of total errors were significantly higher than was indicated by the change in the cohort census due to ICD-10-CM and SNOMED-CT value sets performance (2,515).

Table 10. *CKD: VSAC ICD and SNOMED value set phenotype performance*

CKD Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
ICD-10-CM	Yes	357	220	577	0.242	0.999
	No	1,121	200,219	201,340		
SNOMED-CT	Yes	333	217	550	0.225	0.999
	No	1,145	200,222	201,367		
Total		1,478	200,439	201,917		
(McNamar's test statistic χ2)					**12.250**	0
Critical value 10.827						
df=1, level .001						

For CKD, McNamar's $\chi 2$ (12.250) is greater than the critical value for sensitivity, and the null hypothesis is rejected. McNamar's $\chi 2$ (0) is less than the critical values, and the null hypothesis is accepted for specificity. For CKD, the difference in sensitivity is statistically significant, but specificity is the same for ICD-10 versus SNOMED.

Table 11. *Heart Failure: VSAC ICD and SNOMED value set phenotype performance*

Heart Failure Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
ICD-10-CM	Yes	3,675	0	3,675	0.926	1.000
	No	295	197,943	198,238		
SNOMED-CT	Yes	3,678	35	3,713	0.926	1.000
	No	292	197,908	198,200		
Total		3,970	197,943	201,913		
(McNamar's test statistic χ2)					0	0

Critical value 10.827
df=1, level .001

For heart failure, McNamar's χ2 (0) is greater than the critical value for sensitivity, and the null hypothesis is accepted. The sensitivity and specificity were the same.

3.4 Discussion and Recommendations

The inclusion rules selected to create specific chronic care cohorts are highly complex and dynamic in clinically active electronic health records. The type of terminology selected, such as ICD or SNOMED, may significantly impact cohort attribution.

The evidence shows that in the setting of diabetes, choice of diagnosis terminology does make a statistically significant difference in the phenotype performance whereas this was not shown to be true for heart failure. CKD showed improvement in sensitivity, but not specificity.

A more nuanced interpretation of the results can be made in the context of population health use cases. If the objective is to find and effectively treat patients with uncontrolled Type 2 diabetes, as in eCQM 122 *Percent with HbA1c level > 9.0%,* then errors of false omission from the diabetes cohort will prevent the identification of uncontrolled patients. Error rates can be read directly from the 2x2 contingency tables (Tables 9-11), and ICD had 2575 fewer false omissions for diabetes than SNOMED. ICD also had fewer errors overall than SNOMED for diabetes (7283, 8516), CKD (1341, 1362) and heart failure (295, 327).

There are reasons to believe that this pattern may persist across other disease states. The mapping error rate of ICD is likely to be lower than SNOMED because version ICD-9 was originally developed in 1975 (Moriyama IM, Loy RM, & AHT, 2011), and has decades of use and governance behind it. It's universal, and ingrained in our medical practice. SNOMED-CT's low interrater reliability, would imply that phenotyping results using SNOMED would likely be inconsistent, as the codes used to define a common finding can be (Andrews et al., 2007).

Chapter 4: Aim 2. Measure terminology maintenance impact on SNOMED cohorts

4.1 Introduction

Regulatory updates to address evolving terminology needs must be applied twice yearly in a process that modifies diagnosis terms and maps between ICD and SNOMED. This EHR terminology maintenance can produce secondary changes to cohorts coded in ICD and retrieved by phenotype algorithms using SNOMED CT value sets. This Aim was designed to capture quantitative evidence of the effect of the October 1, 2016 IMO regulatory update on population health cohorts for diabetes, CKD and heart failure. Like Aim 1, the Epic® 2015 Grouper editor utility was used to create the secondary value set of the local diagnosis codes mapped to VSAC value sets thereby externalizing the mapping of local codes to standard diagnosis terminologies. The resolved versions of VSAC value sets were captured before (May 22, 2016) and after the terminology maintenance (November 6, 2016).

4.2 Methodology

The null hypotheses for Aim 2, $H_{0DM,}$ states that for diabetes, CKD and heart disease, the sensitivity and specificity of phenotype algorithms using SNOMED exclusively will be the same before and after terminology maintenance (Equation 2A). The alternative hypothesis, $H_{ADM,}$ states that for each disease, the sensitivity and specificity of phenotype algorithms using SNOMED exclusively will be different before and after terminology maintenance (Equation 2B).

Equation 3. Aim 2 Hypotheses for comparison of SNOMED value sets before and after terminology maintenance

A) Null hypothesis H_{0DM}: $\theta ijk\ Phenotype\{SNOMED, T_1\} = \theta ijk Phenotype\{SNOMED, T_2\}$

B) Alt hypothesis H_{ADM}: $\theta ijk\ Phenotype\{SNOMED, T_1\} \neq \theta ijk Phenotype\{SNOMED, T_2\}$

Where θijk is the true summary measure of the *i*th Disease D_i, $i=\{1,2,\ldots n\}$

and jth Summary Measure M_j, $j=\{1,2\}$ at Time T_k $j=\{$Before, After$\}$ for each phenotype algorithm

Disease D= { DM, CKD, HF }

Summary Measure M= {Sensitivity, Specificity}

The shaded cohort in Figure 12 represents the VSAC SNOMED cohort at Time 2. The same SNOMED and reference standard cohorts described in Aim 1 were also used in Aim 2.

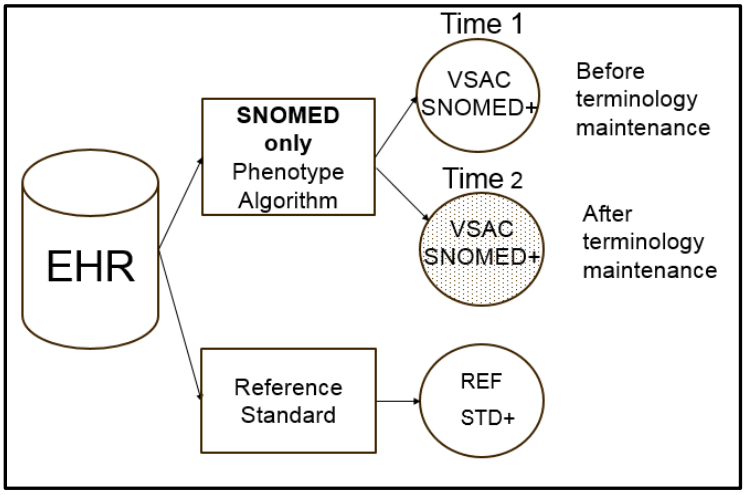
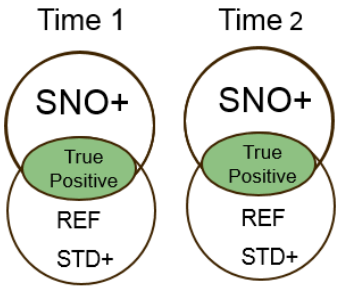


Figure 12. Aim 2 cohorts (before terminology maintenance, after maintenance, and reference standard)

The resulting True Positive, False Positive, True Negative and False Negative values were recorded in 2x2 contingency tables as shown conceptually in Figure 13. These values were used to derive the sensitivity and specificity of the SNOMED diagnosis phenotypes for each disease before and after terminology maintenance.



SNOMED		Reference Standard		Total
		Disease	No Disease	
Time 1	+	**True Positive (TP)**	False Positive (FP)	SNOMED (T1) +
	-	False Negative (FN)	True Negative (TN)	SNOMED (T1) -
Time 2	+	**True Positive (TP)**	False Positive (FP)	SNOMED (T2) +
	-	False Negative (FN)	True Negative (TN)	SNOMED (T2) -

Figure 13. Comparison of VSAC SNOMED cohorts to a reference standard

4.3 Findings

Table 12. Diabetes: *VSAC SNOMED phenotype performance before and after terminology maintenance*

Diabetes Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
SNOMED-CT	Yes	7,471	4,708	12,179	0.662	0.975
Before	No	3,808	185,926	189,734		
SNOMED-CT	Yes	7,708	4,844	12,552	0.683	0.975
After	No	3,571	185,790	189,361		
Total		11,279	190,634	201,913		
(McNamar's test statistic χ2)					**174.596**	**99.049**
Critical value 10.827 df=1, level.001						

McNamar's $\chi 2$ test statistics for diabetes sensitivity (174.596) and specificity (99.049) exceed the critical value of 10.827, and the null hypothesis is rejected for diabetes. Differences in sensitivity and specificity between ICD and SNOMED algorithms are statistically significant ($p = 0.001$). Although the value set had not changed, the local EMR terminology maintenance on October 1, 2016 resulted in increased sensitivity from 0.662 to 0.683 ($p <= 0.001$). The census change in the cohort (237) does not reflect the true magnitude of the change (319) with 278 patients added and 41 patients excluded.

Table 13. *CKD: VSAC SNOMED phenotype performance before and after terminology maintenance*

CKD Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
SNOMED-CT Before	Yes	333	217	550	0.225	0.999
	No	1,145	200,222	201,367		
SNOMED-CT After	Yes	334	217	551	0.226	0.999
	No	1,144	200,222	201,366		
Total		1,478	200,439	201,917		
(McNamar's test statistic $\chi 2$)					0	0

Critical value 10.827
df=1, level .001

McNamar's statistic (0) was less than the critical value (10.827) for both sensitivity and specificity. The null hypothesis is accepted. There was no change for CKD and heart

failure as a result of the terminology maintenance It is important to note that findings may vary by disease state due to the unique features of the lexicon of disease diagnoses.

Table 14. *Heart failure: VSAC SNOMED phenotype performance before and after terminology maintenance*

Heart Failure Value Sets		Reference Standard				
		Yes	No	Total	Sensitivity	Specificity
SNOMED-CT Before	Yes	3,678	35	3,713	0.926	1.000
	No	292	197,908	198,200		
SNOMED-CT After	Yes	3,678	35	3,713	0.926	1.000
	No	292	197,908	198,200		
Total		3,970	197,943	201,913		
(McNamar's test statistic χ2)					0	0

Critical value 10.827
df=1, level .001

McNamar's statistic (0) was less than the critical value (10.827) for both sensitivity and specificity. The null hypothesis is accepted. There was no change for CKD and heart failure as a result of the terminology maintenance It is important to note that findings may vary by disease state due to the unique features of the lexicon of disease diagnoses.

4.4 Discussion and Recommendations

The performance of SNOMED value sets can change over time even if the value set itself remains unchanged. Changes in a cohort census may not reflect the true magnitude of the change because patients added and excluded offset each other. Unexpected changes in

cohort membership require clinical validation to determine if the change is good, bad, or irrelevant. This study demonstrates that periodic EHR maintenance to apply terminology updates may cause discontinuities in the size of chronic disease cohorts. This is not the only cause for unexpected shifts in population health registries census. Technical problems related to the nightly load of registry data marts, EHR application upgrades, and customization activity have been known to affect registry census in dramatic ways. The first step in addressing these problems is to detect them. The development of EHR registry systems that automatically monitor their own performance would remove burden from the IT staff, making it more likely that anomalies are detected.

Recommendation: Best practice for maintenance of EHR registries requires longitudinal monitoring of daily census.

Diagnosis terminology maintenance may cause unpredictable, but potentially large changes in the census of SNOMED-CT cohorts. The development of novel phenotype in SNOMED terminology is easier and less time-consuming because many ICD codes can be represented by a single SNOMED concept. (Lieberman et al., 2003) However, this same characteristic increases the probability of large changes in SNOMED cohorts due to terminology maintenance.

Recommendation: ICD is the preferred terminology for population health cohort algorithms.

Testing can mitigate the impact of terminology maintenance on population health cohorts. Phenotype algorithms using SNOMED value sets should be retested with each terminology update. If significant performance degradation is found, the inclusion rules can be adjusted as necessary, and retested under the new mapping conditions.

Recommendation: Best practice for SNOMED value sets requires retesting of algorithm performance when terminology updates are applied. Adjustments to the code sets can then be applied as needed.

Chapter 5: Aim 3. Recommend methods for improving population health interventions

5.1 Introduction

The development of phenotypes for population health cohorts is embedded within the larger context of deploying technology-aided clinical workflows. A body of work has informed the importance of the socio-technical dimension of EHR implementations, and Aim 3 applies the socio-technical model developed by Sittig and Singh (2010) for studying health information technology in complex adaptive healthcare systems.

5.2 Methodology

The eight dimensions of the socio-technical model all relate in some way to the interaction of patients, the healthcare workers who serve them, and the computer systems that initiate care alerts and track the data collected during the delivery of healthcare. The social perspective, or the human side of this interaction is concerned primarily with the dimensions of People, Workflow and Communication, Internal Organizational Policies, Procedures and Culture and External Rules, Regulations and Pressures. The technical perspective addresses the Hardware and Software Computing Infrastructure, Clinical Content, Human Computer Interface, and System Measurement and Monitoring. Observations across these domains were informed by literature research and experiential knowledge gained over a five year period during which the author was responsible for design, development, implementation and support of a population health registry infrastructure at an academic medical center.

5.3 Findings

The findings are organized by social and technical domains, and informed the practical methodology for improving population health interventions.

5.3.1 Social Domains. *People* refers to anyone who engages in the population health program in some way. *Workflow and Communication* involves the recognition and treatment of chronic disease in population health programs. *Internal Organizational Policies, Procedures and Culture and External Rules* set standards of practice for managing clinical data. All of the domains are impacted by *Regulations and Pressures* driving the adoption of population health care delivery model..

The dimension of *People* as defined in this study includes the humans directly involved in population health programs. Three subcategories were explored: 1) patients; 2) clinicians 3) information technology (IT) workers.

Patients contribute clinical data points used to form disease cohorts, yet each group is comprised of individual patients with unique histories, disease manifestations, and choices. Study participants were weighted toward mature female adults. Two thirds of patients were age 40 or over and 57.5% were female. The prevalence of chronic disease and comorbidity increase with age. Therefore, population health programs can apply lessons learned about communication and treatment preferences for this age group. Recent findings imply that assumptions about the use of communication technology by older adults may be breaking down. Ruppel, Blight, Cherney, and Fylling (2016) found evidence that the text-based format of e-mail might help older adults compensate for hearing impairment of communicative difficulties.

The challenge of chronic disease is to diagnose and treat patients before disease progression permanently damages body systems. Perhaps the greatest opportunities for promoting wellness healthy lifestyles will occur in the 18-40 year old population. This age group is more likely to be comfortable with the online platforms of patient portal. Patient-centered display of clinical markers of disease as well as plans for care help engage the patient in the management of a chronic condition. The portals provide a secure messaging interface to report errors in the data or plans, and patients will likely have a growing role in monitoring and improving their own health care data. To address individual needs, population health intervention triggers must be designed with a personalization feature to turn them off when recommendations have been made in error or an intervention is otherwise contraindicated or refused. Historically, patient advisory boards have had little or no role to date in the design of the patient portal interface for the chronic disease management interfaces, but certainly could be leveraged to provide feedback that would inform improvements in usability and effectiveness.

Clinicians bear the responsibility for applying the clinical diagnoses. Historically, the diagnosis would have been written in free text in a physician note, but this type of information has been a challenge for computers to capture and process. A widespread approach to solving this problem is to have the doctor perform a text search on a diagnosis term, and choose the best match from a drop-down list. The EHR used in this study maps thousands of diagnosis terms between four terminologies, the local diagnosis codes and three standard diagnosis terminologies, in a complex web of many to many relationships. The goal and promise of the terminology vendor is to make it as easy as possible for a provider to find a clinical term (diagnosis) on the list of available terms. A

surplus of synonymous terms may actually impede the selection of the "correct" code, and the mapping adds complexity and fuzziness to the retrieval of these codes. Operating in a time-constrained care delivery system, the clinical collects historical, symptom, exam, laboratory, procedural and genetic information over time to accurately select the right diagnosis. In comparison, the phenotype algorithms being used to group patients for care interventions contain relatively few data points. The patient with chronic disease is also on a dynamic path of disease progression in which age, lifestyle, comorbidities and medication can influence quantitative laboratory measures of a disease state.

Resnik, Niv, Nossal, Kapit, and Toren (2008) made an early study of trade-offs between structured input and unrestricted free text clinical notes, and in fact both formats exist side by side in current EHR. As natural language processing (NLP) systems develop, computer assisted coding in clinical care delivery may develop with similar features to systems currently supporting administrative coding used for billing. However a recent systematic review of NLP systems for capturing unstructured clinical information reports continuing challenges with extraction of temporal information and normalization of concepts to standard terminologies (Kreimeyer et al., 2017).

It should be noted that clinicians engaged in population health programs are often supported by nurses, care managers or health coaches in a team-based care model. Due to the wide variety of licensed and unlicensed healthcare workers and team models, this group of population health practitioners was deemed to be out of scope for this general review.

Information technology (IT) workers recruited into the emerging practice of population health often lack the training and knowledge to navigate complex terminology decision

and implement statistical measurement of phenotypes. There is currently no professional or academic program specific to the development and application of population health interventions, and the content is complex and rapidly changing. The knowledge and practice of population health medicine should be encapsulated in training/certification programs for practitioners. The population health IT analyst performs best in a multidisciplinary team that also includes clinical subject matter experts, clinical operations leaders, data and financial analysts that report to executive leadership.

A challenge in the *Workflow and Communication* dimension relates to the recognition and communication of diagnoses**.** Undiagnosed illness prevents patients from benefiting from treatment, and raises the risk of complications. CDC estimates that there are 30.3 million people with diabetes in the U.S., 7.2 million of those are undiagnosed (Centers for Disease Control and Prevention, 2017). Accurate clinical phenotypes can identify patients with missed diagnosis of chronic disease and inform prognostic prediction models that predict the risk of developing a disease, a comorbid condition, or mortality at a specific point in the future (Hsieh, 2017). New clinical workflows will be necessary to identify, verify, and inform patients that meet a disease phenotype without a corresponding diagnosis. When computer algorithms detect what appears to be a "missed diagnosis" clinical staff need a plan for verifying the diagnosis and sharing this information with the patient. This type of outreach can begin to address the issue of the undiagnosed chronically ill.

The domain of *Organizational policies and procedures* includes the important topic of governance**.** There continues to be sizable local variation in the collection and storage of clinical data. Data governance structures within an organization can lead decision making

related to what data to collect, how to code it, and where to store it. Without health system governance, choices regarding the collection and storage of data may proceed in an ad hoc fashion, complicating the retrieval and analysis of data for population health and quality reporting.

The domain of *External rules, regulations and pressures* encompasses the forces driving the adoption of population health care delivery model. These include technological advances, emphasis on evidence-based care, shift to outpatient care, change to value-based reimbursement and shared risk structures with payers (FitzGerald, 2017).

CMS national quality programs create de facto standards for quality measurement, and incentivize the use of population health interventions. Improvement in the clinical validity of chronic disease value sets by CMS would be an important step toward reducing the national burden of value set development and maintenance. For instance a quality measure about statin therapy for cardiovascular disease should not have a congenital anomaly (not appropriate for statin treatment) included in the diagnosis value sets.

5.3.2 Technical Domain. The technical perspective addresses the *Hardware and Software Computing Infrastructure*, *Clinical Content*, *Human Computer Interface*, and *System Measurement and Monitoring*.

The *Hardware and software* domain is dominated by EHR systems**.** The existence of registry functionality within the EHR system is relatively new. The term Sustainable, Timely, Operational Registries in the EHR (STORE) has been suggested to differentiate this type of registry from traditional registries that are external to the EHR and unable to trigger delivery of clinical care or patient communication (Berkovich, 2016). EHR

registries hold the promise to drive down the cost of care through efficiencies of scale, and are becoming an essential tool in the ambulatory setting for delivering high quality care as defined by evidence-based practice and national quality measurement programs. However, the application tools to develop, validate and manage the clinical interventions in a scientific manner are still evolving. New temporal measures like persistence of abnormal lab results over a 90 day period should come standard in the registry metric calculation toolbox as well as tools to better identify, validate, and quantify and visualize phenotype performance.

The *Clinical content* domain is dependent upon standards. The sharing of diagnosis algorithms and by extension, phenotypes requires adherence to content standards such as LOINC, RXNORM, REAL (Race, Ethnicity, and Language), etc. As was discovered in the implementation of the eMERGE diabetes reference standard, the ability to use a standard does not equate with effective use. The Meaningful Use quality program did effectively require the identification of patient phenotypes by ICD codes, and MACRA MIPS will require reporting of SNOMED codes in 2018. CMS could speed the adoption of standard lab results and medications terminologies by requiring quality measures to be reported using LOINC and RXNORM codes, respectively.

Computer assisted coding (CAC) systems must have a well-defined basis for evaluating their own correctness. To answer the question, "How does the system know it's right?" Resnik offers a generalized formula for sensitivity to quantify the performance of CAC (Jiang, Nossal, & Resnik, 2006).

Equation 4. Resnik's formula: How does the system know it's right?

$$\Pr(\text{choice C is correct} \mid \text{evidence}).$$

To achieve high levels of sensitivity and specificity, previous studies have concluded that ICD-9 codes are not sufficient for identifying patient cohorts (Shivade et al., 2014). In addition to laboratory reports, procedure results, medications, and demographics used in the eMERGE reference standards, phenotypes algorithms may need to include novel data items such as geo codes, social determinants, patient reported data, calculated data such as average blood pressure, and risk scores. Free text evidence in the EHR not routinely recognized by EHR registry inclusion rules could provide additional evidence on which to base phenotyping choices. External data such as public death records may be used in the future to prevent outreach to deceased patients.

The future development of the inclusion algorithms must account for the signal strength of each element. For instance, an encounter diagnosis that appears 24 times in a patient record is more reliable than another that may appear once. Diagnosis codes are also found in Medical History, and billing codes, and free text notes.

The dimension of the *Human-computer interface* has been well studied when applied to the use of an operational EHR application, but less so in the study of the tools used to develop and maintain population health applications. The task of creating and assessing value sets as practiced today is highly complex, and introduces numerous opportunities for human error. The number of codes in diagnosis value sets may be trivial for a computer, but can seem large in human terms. For example the VSAC ICD-10 value set for diabetes had 146 codes and SNOMED had 36 codes. Due to the hierarchical nature concepts, SNOMED CT value sets typically have fewer codes than ICD. Humans manually entering value sets may have a natural bias toward SNOMED CT because it appears to be quicker and easier. This bias does not take into account the retrieval

properties of SNOMED that appear to have lower performance scores and are known to cause unpredictable shifts in some patient cohorts as a result of terminology maintenance. Best practice would dictate that value set editors should display both the computer readable code and the human readable term side by side, but this is not currently a requirement of certified EHR systems. A direct method for importing national value sets into the EHR would also ease the burden of value set creation. Although EHR systems allow a great deal of flexibility in designing an algorithm, the tools to assess the performance of one value set as compared to another or against a reference standard are limited. The addition of statistical analytics and visualizations to phenotyping toolkits would increase the availability of comparative data upon which to make scientifically based phenotype decisions.

The domain of *System measurement and monitoring* is garnering greater attention due in part to patient safety studies on clinical decision support systems (Wright et al., 2017). The choices in the Value Set Authority Center for common diseases exceeds human capacity to select and compare. Searchable ratings indicating a range of value set performance like sensitivity, specificity, false discovery rate, false omission rate in multiple settings should be reported for value sets maintained within its library. VSAC contains value sets, but the development of accurate phenotypes will require multiple terminologies within a single phenotype. These algorithms should be evaluated against standard EHR implementations and published with measures of performance. Similar to national quality metric definitions, national disease phenotypes should be published that include multiple clinical domains like diagnoses, labs, procedures, etc.

Computer Assisted Coding (CAC) systems scans medical records and supports human coders by suggesting ICD billing codes supported by clinical documentation. This model could be used for augmented diagnosis tools that would suggest codes based on clinical evidence, risk factors and local prevalence. Resnik suggests quality measures for CAC systems should address completeness, correctness, and non-redundancy. Resnik discusses a similar process as applied to Computer Assisted Coding.

> When evaluating a system intended to match human expert performance, issues to address include defining test data, selecting performance measures, determining what responses the system should produce, and deciding whether particular levels of performance are "good enough, "(Jiang et al., 2006)

Hripcsak and Heitjan (2002) addressed the problem of assessing the performance of a decision support system when there is no definitive way to know the true state of the patient.. Their study also compiled results in a two-by-two contingency table, and then compared the statistics of observed agreement, specific agreement, and Kappa. Kappa is defined in terms of the observed agreement and agreement expected by chance (Hripcsak & Heitjan, 2002). The specific recommendation for dichotomous data concludes that showing the two-by-two contingency table with its marginal totals is probably as informative as any measure.

Studies comparing overlapping sets may provide formulas that may useful for measuring how much similarity and difference between two phenotype rules run against the same population. The use of Venn Diagrams in visualization may be a cognitive aid to help analysts understand and communicate the difference between the test rule and the reference standard or two versions of the same rule.

5.4 Discussion and Recommendations

The population health care delivery model exists within a socio-technical system, and will need both workflow and technology changes to succeed. The socio-technical model provides a framework from which to study health IT interventions within a complex adaptive health care system (Sittig & Singh, 2010). A Socio-Technical Approach to Population Health Programs is outlined in Table 8. This practical method was synthesized from findings in this study, direct experience and published literature (Shivade et al., 2014).

Table 15. *Socio-Technical Approach to Population Health Programs*

Task	Description
Begin with defined goals	National quality programs are often the starting point for Population Health. Consider available resources to carry out planned interventions
Form the team	Health IT analysts need clinical oversight, operational and executive support
Build a reference standard	Start with a research phenotype or diagnostic tests
Design and test the inclusion rule	Rapid iteration and refinement of the inclusion rule using quantitative phenotype scores will lead to the best inclusion rule
Report phenotype performance including error rates	Report 2x2 contingency table, Sensitivity, Specificity, Error Rates and F score
Implement the population health plan	Find the most efficient and effective outreach methods
Measure success	Collect baseline data, and track progress toward goal

5.4.1 Begin with defined goals. Goals, resource constraints, error limits, and measures of success will all contribute to design decisions when developing a new population health intervention. Many organizations set a goal of "building a registry" or "doing population health" without being specific about the parameters of the project. Regulatory quality measures have pre-defined processes and performance targets. There may also be a strong financial motive to deliver value-based care and earn incentives payments based on quality performance. Population health systems provide a framework to support and measure health interventions, but ultimately the project should demonstrate meaningful results. Health outcomes that matter to patients, providers and payers include reduction in development of new disease, health complications, emergency or hospital visits, and mortality. Health systems may also be financially incentivized or penalized for performance on quality measures. The work queues and intervention alerts must be scaled to the staff available to process the work, therefore resource constraints must be considered in algorithm design. For example, it's counterproductive to refer 1,000 patients to a health coaching resource that can only serve 500. In the setting of capitated payments or value-based care, the size of the patient cohort will inform the costs and resources required to carry out the intervention, whereas medical complexity, multimorbidity and social determinants will impact the intervention design and likelihood of success. The inclusion and exclusion error rates are associated with their own risks and costs, and should be considered when making terminology selections. Resnik discusses a similar process as applied to Computer Assisted Coding.

> When evaluating a system intended to match human expert performance, issues to address include defining test data, selecting performance measures, determining

what responses the system should produce, and deciding whether particular levels of performance are "good enough." (Jiang et al., 2006)

5.4.2 Form the team. Population health is a type of surveillance system used to ensure that patients receive the appropriate standards of care. The requirements for population-level response for patient safety incidents as described by Hibbert et al. (2016) can easily be adapted to population health. A multidisciplinary team that includes clinical subject matter experts, clinical operations leaders, IT and financial analysts, and data scientists will provide the right skillset for a successful population health program. Executive sponsorship and support is generally required within a large organization to direct resources and budget to population health activities. Smaller practices may choose to procure third-party population health services or hire consultants to help them with the task. Patient advisory boards increasingly pivotal in guiding the timing and approach of messaging and patient portal interfaces for chronic disease management.

5.4.3 Build a reference standard. Statistics that compare phenotoype performance to a reference standards are much more informative than those that just compare differences between cohorts. In the latter, positive and negative changes to patient cohorts can offset one another to mask the true number of differences. Validated clinical phenotypes may be found in the medical literature and at PheKB.org a phenotype knowledge based developed the Electronic Medical Records and Genomics (eMERGE) network. Perhaps in the future, open source data sets will support reference standard development.

5.4.4 Design and test the inclusion rule. Organizations committed to the development of a population health program to drive a quality measure to its goal may be tempted to use the quality measure definitions of numerator and denominator as specific

design criteria. However, as demonstrated, the value sets supplied by a quality program may not be optimized for delivery of clinical care. Although national value sets may be a good starting point, clinicians may recommend refinements to remove patients with contraindications or temporary conditions. After diagnosis criteria has been optimized, additional clinical indicators of disease should be iteratively added and tested to drive the performance of the inclusion rule toward the level of the reference standard itself. One may wonder why the reference standard is not implemented directly as the inclusion rule. Some registry platforms are limited to the logical constructs or data types that can be implemented as inclusion rules. For example, the persistence measure for elapsed days between abnormal lab results may not be implemented in the EHR such that it can be utilized as a registry inclusion rule. Work-arounds may exist, but may not be practical with the constraints of budget, skills, and implementation schedules. If the phenotype query captures and reports the data underlying the decision variables the validation time will be significantly reduced.

There is a natural trade-off between sensitivity and specificity which are both measures of intrinsic diagnostic accuracy unaffected by the prevalence of the condition (Zhou, Obuchowski, & McClish, 2011a). Quantitative measurement of the algorithms helps with design decisions. The F score is a convenient single statistic to compare any number of similar algorithms while balancing sensitivity and specificity.

5.4.5 Report phenotype performance including error rates.

From the perspective of cohort development, simplification within and clear separation between cohorts are of high importance. Although phenotyping techniques have improved over the past few years, there is still room for improvement (Shivade et al.,

2014). There is now also a need for EMR phenotype algorithms that can perform well across populations where the patient characteristics may vary (Liao et al., 2015). Too often, the same task is being repeated at multiple institutions (Shivade et al., 2014). Yet, "incomplete reporting has been identified as a major source of avoidable waste in biomedical research," (Bossuyt et al., 2015). The use Standards for Reporting of Diagnostic Accuracy Studies (STARD), or a derivative of those standards for use in population health studies will ensure that essential items are reported in phenotyping studies. Statistics of particular importance for population health are the 2 x 2 contingency table from which one can calculate the False Omission Rate, False Discovery Rate, Sensitivity, Specificity and the F Score.

5.4.6 Implement the Population health plan. Lessons learned from the implementation of Health IT systems can inform the practice of population health. The ten guidelines for HIT Design for Chronic Disease Care provide a useful list to consider. (Unertl, Weinger, Johnson, & Lorenzi, 2009)

5.4.7 Measure your success and share lessons learned.

Healthcare reform in the U.S. focused on providing value for the patient in terms of health outcomes achieved per dollar spent. Measurement and dissemination of health outcomes will become universally mandated. (Porter 2009) Population health programs should be able to demonstrate that they are providing anticipated outcomes. Collect baseline data, and track progress toward the defined goals.

Chapter 6: Synthesis and Summarization

This research confirms that terminology mapping considerations are important in managing population health cohorts. This is a formidable challenge because mapping local data to standard terminologies and between terminologies impacts every EHR system. Additionally, medical terminology systems are complex, and are continuously evolving. Maps between large terminologies and ontologies can introduce phenotype errors, and work is needed to better understand how value sets and mappings shape population health cohorts and affect quality measurement. Although research continues on automating the mapping process, there is still a degree of judgement and imperfection introduced as humans develop terminology maps, apply codes to the patient health record, and retrieve chronic care cohorts.

The terminology vendor, by providing maps from local terms to national standards delivers middleware functions as its own coded terminology. Certified EHRs actually support and cross-map four terminologies (local/vendor supplied, ICD-9, ICD-10, and SNOMED). When a clinician searches an EHR for a diagnosis term that fits a patient's presenting condition, a local code is displayed that maps the clinical term supplied by the vendor to an ICD-10 code. When Brown declared in 2005 that ICD-9 was an obsolete coding system, she was reinforcing the belief that ICD, designed as classification system, could not provide the necessary granularity to capture the clinical language and nuance used by physicians. ICD-9, has not been fully retired, but continues to be applied to

patient records through the terminology mapping process, mainly for research purposes. ICD-11 is under development, and will be finalized in 2018. (World Health Organization) Population health systems must be continually reviewed and maintained to stay in sync with the changes.

6.1 Conclusion and Recommendations

Population Health is growing in importance as a care delivery model as national value-based care programs seek to contain cost and improve outcomes. In the setting of capitated payments and the value-based care paradigm, patient cohort attribution can impact the costs, resources, and interventions. Phenotype inclusion errors raise the risk of patients receiving inappropriate care and exclusion errors raise the risk of missing standard care interventions, and the failure to implement population health programs with little attention to socio-technical dimensions can result in sub-optimal results. CMS is in a unique position to develop and require new scientific methods to assign and evaluate chronic disease cohorts.

In some cases, tools and value sets developed for quality reporting programs are being applied in a clinical context they were not designed to support. As required levels of compliance on some measures approach 100% and sizable incentives and penalties impact individual doctors and health systems alike, there is a new imperative to quantitatively and scientifically evaluate population health cohorts with high levels of precision.

This limited study suggests that ICD should be the preferred terminology for population health cohorts in the absence of data to the contrary. Research has shown diagnosis phenotypes using ICD alone are not sufficient, and diverse data sources are being utilized

to improved phenotype results (Shivade et al., 2014). The set of national value sets for identifying chronic disease cohorts will need to develop into a library of national phenotypes that include ICD values sets, and also look beyond readily available diagnosis codes to include other evidence of disease such as labs, procedures, medications, questionnaires, and medical devices. These phenotypes will include multiple terminologies standards (LOINC, RXNORM, etc), and health systems should be required to produce quality reports using those standards. The use of SNOMED terminology in the MIPS quality program promises to generate valuable data in 2018 and beyond.

Lack of portability and restrictions on data sharing have left the development of clinical cohorts largely isolated within each health system, and it's a herculean task. Within each local instance of an EHR, efficient and accurate data capture and governance is paramount, for this data both drives the delivery of care and collects raw material for Big Data systems. While Big Data may be part of the solution, federated research networks have, by design, added another layer of mapping that masks the operating characteristics of the local EHR. Phenotyping studies in a Big Data network may inform the overall design of phenotype logic, but there is currently no established method for distributing phenotypes across EHR systems. Decentralized local development and common sharing of methods and results may be the only way forward in the near future. Common standards for reporting methods and de-identified aggregated results would be required for this approach to be effective. STARD (Standards for Reporting Diagnostic Accuracy), which is currently used for diagnostic studies provides a relevant model.

Although the focus on population health programs is often centered on the technology, the social aspects of system development implementation and measurement equally

apply. Local health systems need multidisciplinary teams who are trained to improve phenotype testing and refine diagnostic algorithms using quantitative scientific methods. Data governance bodies within the health system are needed to inform decisions on what data to collect, how to code it, and where to store it. As national, state, professional, and local quality programs continue to mandate the reporting of quality measurements, it is a shared responsibility of the medical community to ensure that the data, itself, is high quality.

Terminology mappings within an EHR can be very difficult to discern by clinicians and application analysts. Population health platforms, whether they exist within an EHR or a third party provider should deliver tools for monitoring and reporting fluctuations in registry census. Without this feature, unexpected changes to patient cohorts may be go undetected, but should be evaluated by a clinician. A summary of these recommendations may be found in Appendix E. Recommendations for Quality and Population Health Programs.

6.2 Limitations

This study was limited to one organization, but the inclusion of the entire population of nearly 200,000 patients enabled the calculation of the phenotype performance with high precision ($p=0.001$). Nevertheless, it is the methodology and practice of population health phenotyping that this work seeks to inform. Only one EHR platform was studied, but the platform is one of the top six, and used by 55% of the customer base selected to implement population health systems within their local EHR. According to HIMSS, "less than a third U.S. hospitals surveyed are using a solution from their vendor for population health," (FitzGerald, 2017). The single terminology product included in the study is

incorporated into several of the leading EHR systems, and is nearly universal within the industry. Alternative terminology resources are U.S. National Library of Medicine (NLM) for SNOMED CT, and the Department of Health and Human Services for ICD-10. Apelon® also offers an open source terminology product that supports HL7's FHIR® Terminology Service. The value sets were drawn from a single source, the Value Set Authority Center, the repository for official versions of diagnosis value sets for regulatory quality programs such as Meaningful Use and Clinical Quality Measure. It is the only source of publicly available SNOMED value sets.

Only three disease states were considered. Nevertheless, this study clearly demonstrates that each disease condition has its own unique challenges and characteristics. Diabetes is the sentinel disease in the phenotypes studied. Numerous studies have linked diabetes to CKD and heart failure, including United Kingdom Prospective Diabetes Study. UKPDS developed a model to estimate the lifetime health outcomes of patients with Type 2 Diabetes based on their likelihood of developing renal failure, heart failure, ischemic heart disease, myocardial infarction, stroke, amputation, or blindness (Clarke, 2013). The prevention and treatment of CKD is a key priority of the Healthy People 2020 initiative coordinated by the U.S. Department of Health and Human Services. To that end Healthy People 2020 goals include increased testing of microalbumin levels in persons diagnosed with diabetes, and increased testing of serum creatinine, lipid, and microalbumin in persons with CKD (United States Renal Data System, 2017). CKD has been targeted as a model for improving chronic disease through electronic health records because the disease is common, and objective laboratory data is used for diagnosis and monitoring of disease progression (Drawz et al., 2015) (Navaneethan et al., 2013). The heart failure

algorithm was only designed to detect systolic failure, which lowers the LVEF. Diastolic failure was not included in the phenotype.

Evidence in the reference standard was based on a limited dataset of discrete data. For example, diagnosis information was evaluated in problem lists and ambulatory encounter diagnoses. More evidence of disease state could be found in free text notes, scanned PDF files, external data records, and other discrete locations for diagnosis codes such as inpatient diagnoses or billing records. Information on patient deaths and change of health care provider impact cohort size as well, but this information is not routinely entered into the clinical record because these patients no longer have an active patient relationship within the health system. More data, however, does not necessarily mean better data. Each new data source for each disease state needs to be evaluated at the local level in isolation to avoid adding data that will introduce more uncertainty and noise than valuable insight. Associations between data could also be used as to improve phenotype performance. For example, in CAC systems, the computer can use a crosswalk table to determine which ICD diagnosis codes and CPT procedure codes can be used together (Jiang et al., 2006). A recent review of published phenotype studies show that probabilistic methods and NLP techniques have been gaining popularity as compared to rule-based systems (Shivade et al., 2014).

6.3 Future Work

As value-based care becomes more prominent both as a care paradigm and a reimbursement philosophy, the number of organizations engaged in population health interventions is on a steep growth curve. A HIMSS Analytics survey of 104 IT leaders in

U.S. hospitals found that the number of organizations with population health programs has grown from 67% to 76 % from 2015-2016 (FitzGerald, 2017). Key challenges reported focused on technology, data and resources. To achieve improvements in analytics, performance measurement, and care coordination, population health tools must continue to develop in sophistication, while also becoming easier to use. There is a knowledge gap about what needs to be done, and proper implementation and roll out of programs (FitzGerald, 2017).

As health care providers strive deliver quality care and meet measurement targets, dynamic registries in the EHR that automatically refresh in near real-time are fundamental building blocks for the delivery of population health interventions.

> "Population Health adopters averaged lower acute occupancy rates than non-adopters. Hospitals with 501+ beds that adopted population health average 36% lower occupancy rates than non-adopters," (FitzGerald, 2017)

Well implemented population health programs will underpin the success of the transition to value-based care, and ultimately to the reduction of the burden of chronic disease. However the current trajectory is challenged by high levels of complexity with regard to patient diagnosis. To meet the demands of quality measures and regulatory reporting, a move away from complex custom development to scalable, agile methods could provide the best path forward. Kannan et al. (2017) were able to accelerate the deployment of population registries using a finite set of core principles and a re-usable technology toolkit. Although the development of accurate population health phenotypes will

continue, perhaps an improved short-term solution would be to build population health interventions based on a common treatment *plans* instead of a specific diagnosis.

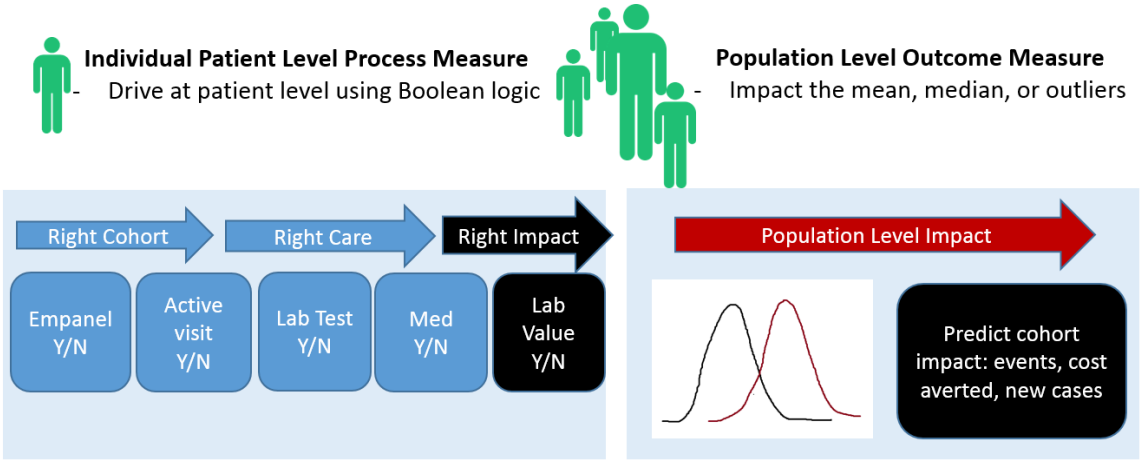


Figure 14. Process drives outcome: transformation of individual care to population impact (Sitapati, A. M., Berkovich B., 2017)

Recurring process cycles applied individuals are at the core of the population health model as shown in Figure 14 Process drives outcome: transformation of individual care to population impact. Chronic care treatment plans frequently follow similar patterns of clinic visits, lab tests, and medication adjustments. When the right cohort receives the right care, and achieves the right impact, the population effects reduce adverse events, costs, and newly diagnosed cases. Population health interventions of the future may rely more heavily on computer-based recommendations to assist clinicians in optimizing complex long-term chronic care planning. Yet there are ample opportunities for population health programs of today to effectively and efficiently deliver personalized plans to benefit of the whole patient population.

References

Agarwal, V., Podchiyska, T., Banda, J. M., Goel, V., Leung, T. I., Minty, E. P., . . . Shah, N. H. (2016). Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc, 23*(6), 1166-1173. doi:10.1093/jamia/ocw028

Altavela, J. L., Dorward, K. M., Sorrento, T. A., Diehl, K. M., & Wyman, C. A. (2017). Population health management: An independent physician organization approach. *American Journal of Health-System Pharmacy, 74*(18), 1477-1485. doi:10.2146/ajhp161009

Anderson, R. N., Miniño, A. M., Hoyert, D. L., & Rosenberg, H. M. (2001). Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. *National vital statistics reports, 49*(2), 1-32.

Andrade, S. E., Moore Simas, T. A., Boudreau, D., Raebel, M. A., Toh, S., Syat, B., . . . Platt, R. (2011). Validation of algorithms to ascertain clinical conditions and medical procedures used during pregnancy. *Pharmacoepidemiol Drug Saf, 20*(11), 1168-1176. doi:10.1002/pds.2217

Andrews, J. E., Richesson, R. L., & Krischer, J. (2007). Variation of SNOMED CT coding of clinical research concepts among coding experts. *J Am Med Inform Assoc, 14*(4), 497-506. doi:10.1197/jamia.M2372

Baker, D. W., Persell, S. D., Thompson, J. A., Soman, N. S., Burgner, K. M., Liss, D., &

Kmetik, K. S. (2007). Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med, 146*(4), 270-277.

Berkovich, B., Sitapati, AM. (2016). Presentation: Development of a Health Registry Ontology: American Medical Informatics Association

Bielinski, S. J., Olson, J. E., Pathak, J., Weinshilboum, R. M., Wang, L., Lyke, K. J., . . . Kullo, I. J. (2014). Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. *Mayo Clin Proc, 89*(1), 25-33. doi:10.1016/j.mayocp.2013.10.021

Bodenreider, O., Nguyen, D., Chiang, P., Chuang, P., Madden, M., Winnenburg, R., . . . D'Souza, I. (2013). The NLM value set authority center. *Stud Health Technol Inform, 192*, 1224.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . Cohen, J. F. (2015). STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Clin Chem, 61*(12), 1446-1452. doi:10.1373/clinchem.2015.246280

Bowman, S. E. (2005). Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems. *Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems/AHIMA, American Health Information Management Association*.

Boyd, A. D., Yang, Y. M., Li, J. R., Kenost, C., Burton, M. D., Becker, B., & Lussier, Y. A. (2015). Challenges and remediation for Patient Safety Indicators in the transition to ICD-10-CM. *Journal of the American Medical Informatics*

Association, 22(1), 19-28. doi:10.1136/amiajnl-2013-002491

Brieler, J. A., Lustman, P. J., Scherrer, J. F., Salas, J., & Schneider, F. D. (2016). Antidepressant medication use and glycaemic control in co-morbid type 2 diabetes and depression. *Fam Pract, 33*(1), 30-36. doi:10.1093/fampra/cmv100

Broberg, C., McLarry, J., Mitchell, J., Winter, C., Doberne, J., Woods, P., . . . Weiss, J. (2015). Accuracy of administrative data for detection and categorization of adult congenital heart disease patients from an electronic medical record. *Pediatr Cardiol, 36*(4), 719-725. doi:10.1007/s00246-014-1068-2

Brown, P. J., & Sonksen, P. (2000). Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *J Am Med Inform Assoc, 7*(4), 392-403. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC61443/pdf/0070392.pdf

Centers for Disease Control and Prevention. (2015). *Diabetes Report Card 2014*. Retrieved from Atlanta, GA:

Centers for Disease Control and Prevention. (2017, July 17, 2017). National Diabetes Statistics Report. 2017. Retrieved from https://www.cdc.gov/diabetes/data/statistics/statistics-report.html

Centers for Medicare & Medicaid Services (CMS). (2016). *CMS Financial Report FY 2016*. (952016). Baltimore, MD Retrieved from www.cms.gov/CFOReport.

Cimino, J. J., Clayton, P. D., Hripcsak, G., & Johnson, S. B. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association, 1*(1), 35.

Cipparone, C. W., Withiam-Leitch, M., Kimminau, K. S., Fox, C. H., Singh, R., & Kahn,

L. (2015). Inaccuracy of ICD-9 Codes for Chronic Kidney Disease: A Study from Two Practice-based Research Networks (PBRNs). *J Am Board Fam Med, 28*(5), 678-682. doi:10.3122/jabfm.2015.05.140136

Clarke, L. (2013, 20-23 Oct. 2013). *Using process modeling and analysis techniques to reduce errors in healthcare.* Paper presented at the Formal Methods in Computer-Aided Design (FMCAD), 2013.

Code of Federal Regulations (annual edition). (2015). *Title 45: Public Welfare. Subpart J: Code Sets 45 CFR 162.1002 Medical data code sets*. U.S. Government Publishing Office: U.S. Government Publishing Office Retrieved from https://www.gpo.gov/fdsys/search/searchresults.action;jsessionid=M8LSPlRrf_YMsR1x6lXKW5eej9nrHp4HPCorPS_1bzjUqZFdEwWf!707445283!513118621?st=45+CFR+162.1002.

cohort. *Stedman's Medical Dictionary* (2016 ed.). Baltimore, Maryland: Lippincott Williams & Wilkens A Wolters Kluwer Health Company.

de Keizer, N. F., Abu-Hanna, A., & Zwetsloot-Schonk, J. H. M. (2000). Understanding terminological systems I: Terminology and typology. *Methods Inf Med, 39*(1), 16-21. Retrieved from <Go to ISI>://WOS:000086308800004

Drawz, P. E., Archdeacon, P., McDonald, C. J., Powe, N. R., Smith, K. A., Norton, J., . . . Narva, A. (2015). CKD as a Model for Improving Chronic Disease Care through Electronic Health Records. *Clin J Am Soc Nephrol, 10*(8), 1488-1499. doi:10.2215/cjn.00940115

Fenton, S. H., & Benigni, M. S. (2014). Projected impact of the ICD-10-CM/PCS conversion on longitudinal data and the Joint Commission Core Measures.

Perspect Health Inf Manag, 11, 1g. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4142515/pdf/phim0011-0001g.pdf

Ferris, M., Shoham, D. A., Pierre-Louis, M., Mandhelker, L., Detwiler, R. K., & Kshirsagar, A. V. (2009). High prevalence of unlabeled chronic kidney disease among inpatients at a tertiary-care hospital. *Am J Med Sci, 337*(2), 93-97. doi:10.1097/MAJ.0b013e318181288e

FitzGerald, B. (2017). *A Look at Trends in Population Health* Presentation Slides. HIMSS 17. HIMSS Analytics.

Floyd, J. S., Blondon, M., Moore, K. P., Boyko, E. J., & Smith, N. L. (2016). Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of Veterans with diabetes. *Pharmacoepidemiol Drug Saf, 25*(4), 467-471. doi:10.1002/pds.3921

Foley, M., Hall, C., Perron, K., & D Andrea, R. (2007). Translation, Please: Mapping Translates Clinical Data between the Many Languages That Document It. *JOURNAL-AHIMA, 78*(2), 34. Retrieved from http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_033474.hcsp?dDocName=bok1

files/1546/doc.html

Fort, D., Wilcox, A. B., & Weng, C. (2014). Could Patient Self-reported Health Data Complement EHR for Phenotyping? *AMIA Annu Symp Proc, 2014*, 1738-1747. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419899/pdf/1986293.pdf

Fung, K. W., & Xu, J. (2012). Synergism between the mapping projects from SNOMED CT to ICD-10 and ICD-10-CM. *AMIA Annu Symp Proc, 2012*, 218-227.

Garvin, J. H., Elkin, P. L., Shen, S., Brown, S., Trusko, B., Wang, E., . . . Speroff, T. (2013). Automated quality measurement in Department of the Veterans Affairs discharge instructions for patients with congestive heart failure. *J Healthc Qual, 35*(4), 16-24. doi:10.1111/j.1945-1474.2011.195.x

Giannangelo, K., & Fenton, S. H. (2008). SNOMED CT survey: an assessment of implementation in EMR/EHR applications. *Perspect Health Inf Manag, 5*(7).

Hibbert, P. D., Healey, F., Lamont, T., Marela, W. M., Warner, B., & Runciman, W. B. (2016). Patient safety's missing link: using clinical expertise to recognize, respond to and reduce risks at a population level. *Int J Qual Health Care, 28*(1), 114-121. doi:10.1093/intqhc/mzv091

Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: History and Context. *American Journal of Neuroradiology, 37*(4), 596-599. doi:10.3174/ajnr.A4696

Hoang, A., Shen, C. Y., Zheng, J., Taylor, S., Groh, W. J., Rosenman, M., . . . Chen, P. S. (2014). Utilization rates of implantable cardioverter-defibrillators for primary prevention of sudden cardiac death: A 2012 calculation for a midwestern health referral region. *Heart Rhythm, 11*(5), 849-855. doi:10.1016/j.hrthm.2014.02.019

Hripcsak, G., & Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. *J Biomed Inform, 35*(2), 99-110.

Hsieh, K. L. (2017). *Building a T2DM prognostic prediction model by incorporating temporal representation*. Research proposal for Advancement to Candidacy for

PhD, University of Texas, Health Science Center at Houston, School of Biomedical Informatics.

Huopaniemi, I., Nadkarni, G., Nadukuru, R., Lotay, V., Ellis, S., Gottesman, O., & Bottinger, E. P. (2014). Disease progression subtype discovery from longitudinal EMR data with a majority of missing values and unknown initial time points. *AMIA Annu Symp Proc, 2014*, 709-718. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419979/pdf/1986499.pdf

Hussain, S., Sun, H., Erturkmen, G. B. L., Yuksel, M., Mead, C., Gray, A. J., . . . Batchelor, C. R. (2014). A justification-based semantic framework for representing, evaluating and utilizing terminology mappings. *Context*, 98-113. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.9002&rep=rep1&type=pdf

Hyde, L., Rihanek, T., Santana-Johnson, T., & al, e. (2013). *Data Mapping and its Impact on Data Integrity*. Retrieved from

IBM. (2016). SPSS Statistics for Windows, Version 24.0.0.0 64 bit. from IBM Corp

Jiang, Y., Nossal, M., & Resnik, P. (2006). How Does the System Know It's Right? Automated Confidence Assessment for Compliant Coding *J Immunol*: Perspectives in Health Information Management, CAC Proceedings.

Jolly, S. E., Navaneethan, S. D., Schold, J. D., Arrigain, S., Sharp, J. W., Jain, A. K., . . . Nally, J. V. (2014). Chronic kidney disease in an electronic health record problem list: quality of care, ESRD, and mortality. *Am J Nephrol, 39*(4), 288-296. doi:10.1159/000360306

Kannan, V., Fish, J. S., Mutz, J. M., Carrington, A. R., Lai, K., Davis, L. S., . . . Willett, D. L. (2017). Rapid Development of Specialty Population Registries and Quality Measures from Electronic Health Record Data*. An Agile Framework. *Methods Inf Med, 56*(99), e74-e83. doi:10.3414/ME16-02-0031

Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., . . . Lowe, W. L. (2012). Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc, 19*(2), 212-218. doi:10.1136/amiajnl-2011-000439

Kim, T. Y. (2016). Automating lexical cross-mapping of ICNP to SNOMED CT. *Inform Health Soc Care, 41*(1), 64-77. doi:10.3109/17538157.2014.948173

Kleinberg, S., & Elhadad, N. (2013). Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. *AMIA Annu Symp Proc, 2013*, 786-795. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900216/pdf/amia_2013_symposium_786.pdf

Klompas, M., Eggleston, E., McVetta, J., Lazarus, R., Li, L., & Platt, R. (2013). Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care, 36*(4), 914-921. doi:10.2337/dc12-0964

Kottke, T. E., & Baechler, C. J. (2013). An algorithm that identifies coronary and heart failure events in the electronic health record. *Prev Chronic Dis, 10*, E29. doi:10.5888/pcd10.120097

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., . . . Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform, 73*, 14-29. doi:10.1016/j.jbi.2017.07.012

Lee, D., Cornet, R., Lau, F., & De Keizer, N. (2013). A survey of SNOMED CT implementations. *J Biomed Inform, 46*(1), 87-96. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046412001530

files/1365/S1532046412001530.html

http://ac.els-cdn.com/S1532046412001530/1-s2.0-S1532046412001530-main.pdf?_tid=41b4de7c-1b01-11e6-9bfa-00000aacb362&acdnat=1463360469_0cb46d80eb2fa945d951458069ab91bf

Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. *The Lancet, 379*(9811), 165-180.

Liao, K. P., Ananthakrishnan, A. N., Kumar, V., Xia, Z., Cagan, A., Gainer, V. S., . . . Cai, T. (2015). Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. *PLoS One, 10*(8), e0136651. doi:10.1371/journal.pone.0136651

Liaw, S. T., Taggart, J., Yu, H., de Lusignan, S., Kuziemsky, C., & Hayen, A. (2014). Integrating electronic health record information to support integrated care: practical application of ontologies to improve the accuracy of diabetes disease registers. *J Biomed Inform, 52*, 364-372. doi:10.1016/j.jbi.2014.07.016

Lieberman, M. I., Ricciardi, T. N., Masarie, F. E., & Spackman, K. A. (2003). The use of SNOMED CT simplifies querying of a clinical data warehouse. *AMIA Annu Symp*

Proc, 910.

Lyon, R. K., & Slawson, J. (2011). An organized approach to chronic disease care. *Fam Pract Manag, 18*(3), 27-31.

McClatchey, S. (2001). Disease management as a performance improvement strategy. *Top Health Inf Manage, 22*(2), 15-23.

Mcleod, A., Keefe, Alyssa, Kemp, Amber. (2016, 1/4/2016). CMS Approves California's Medi-Cal 2020 Demonstration Waiver. Retrieved from www.calhospital.org

Meyers, J. L., Candrilli, S. D., & Kovacs, B. (2011). Type 2 diabetes mellitus and renal impairment in a large outpatient electronic medical records database: rates of diagnosis and antihyperglycemic medication dose adjustment. *Postgrad Med, 123*(3), 133-143. doi:10.3810/pgm.2011.05.2291

Monsen, K. A., Finn, R. S., Fleming, T. E., Garner, E. J., LaValla, A. J., & Riemer, J. G. (2014). Rigor in electronic health record knowledge representation: lessons learned from a SNOMED CT clinical content encoding exercise. *Informatics for Health and Social Care*, 1-15. Retrieved from http://www.tandfonline.com/doi/abs/10.3109/17538157.2014.965302

files/1592/17538157.2014.html

Moriyama IM, Loy RM, & AHT, R.-S. (2011). *History of the statistical classification of diseases and causes of death. Rosenberg HM, Hoyert DL eds*. Hyattsville, MD: National Center for Health Statistics.

Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, . . . Neumar RW, N. G., Palaniappan L, Pandey DK, Reeves MJ, Rodriguez CJ, Rosamond W, Sorlie PD, Stein J, Towfighi A, Turan TN, Virani SS, Woo D, Yeh RW, Turner

MB; on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee,. (2016). Heart Disease and Stroke Statistics—2016 Update: a report from the American Heart Association. *Circulation, 133*, e38-e360.

Murff, H. J., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., . . . Speroff, T. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA, 306*(8), 848-855. doi:10.1001/jama.2011.1204

Nadkarni, G. N., Gottesman, O., Linneman, J. G., Chase, H., Berg, R. L., Farouk, S., . . . Bottinger, E. P. (2014). Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc, 2014*, 907-916. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419875/pdf/1986475.pdf

Nag, S. S., Daniel, G. W., Bullano, M. F., Kamal-Bahl, S., Sajjan, S. G., Hu, H., & Alexander, C. (2007). LDL-C goal attainment among patients newly diagnosed with coronary heart disease or diabetes in a commercial HMO. *J Manag Care Pharm, 13*(8), 652-663. doi:10.18553/jmcp.2007.13.8.652

Navaneethan, S. D., Jolly, S. E., Schold, J. D., Arrigain, S., Saupe, W., Sharp, J., . . . Nally, J. V., Jr. (2011). Development and validation of an electronic health record-based chronic kidney disease registry. *Clin J Am Soc Nephrol, 6*(1), 40-49. doi:10.2215/cjn.04230510

Navaneethan, S. D., Jolly, S. E., Sharp, J., Jain, A., Schold, J. D., Schreiber, M. J., Jr., & Nally, J. V., Jr. (2013). Electronic health records: a new tool to combat chronic

kidney disease? *Clin Nephrol, 79*(3), 175-183. doi:10.5414/cn107757

Pacheco, J. A., Thompson, W. (2012). Type 2 Diabetes Mellitus Phenotype Algorithm. *PheKB.* Retrieved from https://phekb.org/phenotype/type-2-diabetes-mellitus

Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. C., Chen, P. J., . . . Chute, C. G. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J Am Med Inform Assoc, 20*(e2), e341-348. doi:10.1136/amiajnl-2013-001939

Pathak, J., Kiefer, R. C., Bielinski, S. J., & Chute, C. G. (2012). Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. *AMIA Annu Symp Proc, 2012*, 699-708. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540447/pdf/amia_2012_symp_0699.pdf

Porter , M. E. (2009). A Strategy for Health Care Reform — Toward a Value-Based System. *New England Journal of Medicine, 361*(2), 109-112. doi:10.1056/NEJMp0904131

Randorff Hojen, A., & Kuropatwa, R. (2014). SNOMED CT Starter Guide: International Health Standards Development Organisation (IHTSDO).

Rathmann W, & Giani G. (2004). Global Prevalence of Diabetes: Estimates for the Year 2000 and Projections for 2030. *Diabetes Care, 27*(10), 2568-2569.

Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E., . . . Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform, 45*(4), 763-771. doi:10.1016/j.jbi.2012.01.009

Reich, C., Ryan, P. B., Stang, P. E., & Rocca, M. (2012). Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform, 45*(4), 689-696. doi:10.1016/j.jbi.2012.05.002

Resnik, P., Niv, M., Nossal, M., Kapit, A., & Toren, R. (2008). *Communication of Clinically Relevant Information in Electronic Health Records: A Comparison between Structured Data and Unrestricted Physician Language.* Paper presented at the Perspectives in Health Information Management, CAC Proceedings.

Rubbo, B., Fitzpatrick, N. K., Denaxas, S., Daskalopoulou, M., Yu, N., Patel, R. S., & Hemingway, H. (2015). Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *Int J Cardiol, 187*, 705-711. doi:10.1016/j.ijcard.2015.03.075

Ruppel, E. K., Blight, M. G., Cherney, M. R., & Fylling, S. Q. (2016). An Exploratory Investigation of Communication Technologies to Alleviate Communicative Difficulties and Depression in Older Adults. *J Aging Health, 28*(4), 600-620. doi:10.1177/0898264315599942

Saitwal, H., Qing, D., Jones, S., Bernstam, E. V., Chute, C. G., & Johnson, T. R. (2012). Cross-terminology mapping challenges: a demonstration using medication terminological systems. *J Biomed Inform, 45*(4), 613-625. doi:10.1016/j.jbi.2012.06.005

Schildcrout, J. S., Basford, M. A., Pulley, J. M., Masys, D. R., Roden, D. M., Wang, D., . . . Denny, J. C. (2010). An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records. *J Biomed*

Inform, 43(6), 914-923. doi:10.1016/j.jbi.2010.07.011

Schroeder, E. B., Powers, J. D., O'Connor, P. J., Nichols, G. A., Xu, S., Desai, J. R., . . . Steiner, J. F. (2015). Prevalence of chronic kidney disease among individuals with diabetes in the SUPREME-DM Project, 2005-2011. *J Diabetes Complications, 29*(5), 637-643. doi:10.1016/j.jdiacomp.2015.04.007

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc, 21*(2), 221-230. doi:10.1136/amiajnl-2013-001935

Sitapati, A., Berkovich, B. (2016). Presentation: Use of a Local Learning Healthcare System to Improve the Health of Patient and Community: American Medical Informatics Association iHealth 2016.

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care, 19 Suppl 3*, i68-74. doi:10.1136/qshc.2010.042085

So, L., Evans, D., & Quan, H. (2006). ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. *BMC Health Serv Res, 6*, 161. doi:10.1186/1472-6963-6-161

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association, 17*(6), 646-651. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000748/pdf/amiajnl1024.pdf

U.S. National Library of Medicine. Value Set Authority Center. Retrieved from https://vsac.nlm.nih.gov/

U.S. National Library of Medicine. (2011, 7/25/2016). US Extension to SNOMED CT®. Retrieved from https://www.nlm.nih.gov/research/umls/Snomed/us_extension.html

Udris, E. M., Au, D. H., McDonell, M. B., Chen, L., Martin, D. C., Tierney, W. M., & Fihn, S. D. (2001). Comparing methods to identify general internal medicine clinic patients with chronic heart failure. *Am Heart J, 142*(6), 1003-1009. doi:10.1067/mhj.2001.119130

Unertl, K. M., Weinger, M. B., Johnson, K. B., & Lorenzi, N. M. (2009). Describing and modeling workflow and information flow in chronic disease care. *J Am Med Inform Assoc, 16*(6), 826-836. doi:10.1197/jamia.M3000

United States Renal Data System. (2016). *2016 USRDS annual data report: Epidemiology of kidney disease in the United States.* Bethesda, MD Retrieved from https://www.usrds.org/2016/view/Default.aspx.

United States Renal Data System. (2017). *2017 USRDS annual data report: Epidemiology of kidney disease in the United States*. Bethesda, MD Retrieved from https://www.usrds.org/2017/view/Default.aspx.

Wei, W. Q., Leibson, C. L., Ransom, J. E., Kho, A. N., Caraballo, P. J., Chai, H. S., . . . Chute, C. G. (2012). Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc, 19*(2), 219-224. doi:10.1136/amiajnl-2011-000597

Wei, W. Q., Leibson, C. L., Ransom, J. E., Kho, A. N., & Chute, C. G. (2013). The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform, 82*(4), 239-247. doi:10.1016/j.ijmedinf.2012.05.015

Wei, W. Q., Teixeira, P. L., Mo, H., Cronin, R. M., Warner, J. L., & Denny, J. C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc, 23*(e1), e20-27. doi:10.1093/jamia/ocv130

Wilke, R. A., Berg, R. L., Peissig, P., Kitchner, T., Sijercic, B., McCarty, C. A., & McCarty, D. J. (2007). Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res, 5*(1), 1-7. doi:10.3121/cmr.2007.726

World Health Organization. International Classification of Diseases (ICD) Information Sheet. Retrieved from http://www.who.int/classifications/icd/factsheet/en/

Wright, A., Ai, A., Ash, J., Wiesen, J. F., Hickman, T.-T. T., Aaron, S., . . . Sittig, D. F. (2017). Clinical decision support alert malfunctions: analysis and empirically derived taxonomy. *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocx106

Zheng, J., Yarzebski, J., Ramesh, B. P., Goldberg, R. J., & Yu, H. (2014). Automatically Detecting Acute Myocardial Infarction Events from EHR Text: A Preliminary Study. *AMIA Annu Symp Proc, 2014*, 1286-1293. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419972/pdf/1986665.pdf

Zhong, V. W., Obeid, J. S., Craig, J. B., Pfaff, E. R., Thomas, J., Jaacks, L. M., . . .

Mayer-Davis, E. J. (2016). An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. *J Am Med Inform Assoc*. doi:10.1093/jamia/ocv207

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2011a). Chapter 2 Measures of Diagnostic Accuracy *Statistical Methods in Diagnostic Medicine* (pp. 13-55): John Wiley & Sons, Inc.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2011b). Chapter 5 Comparing the Accuracy of Two Diagnostic Tests *Statistical Methods in Diagnostic Medicine* (pp. 165-192): John Wiley & Sons, Inc.

Zwinderman, A. H., Glas, A. S., Bossuyt, P. M., Florie, J., Bipat, S., & Stoker, J. (2008). Statistical models for quantifying diagnostic accuracy with multiple lesions per patient. *Biostatistics*, *9*(3), 513-522. doi:10.1093/biostatistics/kxm052

Appendix A: Literature Review Search Terms

Table 16. *Literature search terms and results*

	Search	Intent	Articles returned
1	(("Heart Diseases"(Zheng, Yarzebski, Ramesh, Goldberg, & Yu) OR heart disease(Zwinderman et al.) OR cardiovascular disease[tw]) OR ("Diabetes Mellitus"[Mesh] OR "Diabetes Insipidus"[Mesh] OR Diabetes[tw]) OR ("Renal Insufficiency, Chronic"[Mesh] OR chronic kidney disease[tw])	To identify a set of articles addressing the treatment of Diabetes, Chronic Kidney Disease, or Heart Disease	1,594,280
2	"Quality Indicators, Health Care"[Mesh] OR "value sets"[Title/Abstract] OR Quality Indicators[Title/Abstract]	To identify a set of articles addressing quality indicators for health care or "value sets"	18,103
3	"Vocabulary, Controlled"[Mesh] OR "Systematized Nomenclature of Medicine" OR "SNOMED" OR "International Classification of Diseases" OR Phenotype OR phenotyping	To identify a set of articles about use of a controlled vocabulary like SNOMED or the International Classification of Diseases for phenotyping	458,707

	Search	Intent	Articles returned
4	("Electronic Health Records"[Mesh] OR "Electronic Health Record"[Text Word] OR Electronic Medical Record [Text Word] OR "EPIC"[Text Word] OR "Medical Records Systems, Computerized"[Mesh])	To identify a set of articles about Electronic Health Records or alternate terms for EHRs	35,930
5	("Vocabulary, Controlled"[Mesh] OR "Phenotype"[Mesh] OR "Systematized Nomenclature of Medicine" OR "SNOMED" OR "International Classification of Diseases" OR "ICD" OR Phenotype OR phenotyping)	To identify a set of articles about use of a controlled vocabulary or Phenotype such as SNOMED or ICD	477,503
6	#1 AND #2 AND #3	Articles discussing Quality Indicators or Value Sets for Phenotyping in SNOMED or International Classification of Disease in the setting of Heart Disease, Diabetes, or Chronic Kidney Disease	34 publications
7	#1 AND #4 AND #5	Articles discussing Electronic Health Records using or phenotyping or controlled vocabularies like SNOMED or ICD the setting of Heart Disease, Diabetes, or Chronic Kidney Disease	193 publications

Appendix B: Electronic Phenotyping Evaluation Methods

Table 17. *Methods used to evaluate electronic phenotypes (from literature search)*

Statistical Method	Coded Terminology n=54	Text Mining n=11	Hybrid n=5
Sensitivity 25	Anderson 2016 Asao 2015 Bagheri 2009 Baker 2007 Broberg 2015 Coleman 2015 Ferris 2009 Floyd 2016 Fort 2014 Garvin 2013 Kleinberg 2013 Lawrence 2013 Liaw 2011 Navaneethan 2011 Onofrei 2004 Rosenman 2013 So 2006 Thiru 2003 AMIA Symp Thiru 2009 Udris 2001 Wei… Chute 2011 Zhong 2015		Brown & Sonksen, 2000 Garvin 2013 Liao 2015 Murff 2011
2x2 contingency table 9	Garvin 2013 Kleinberg 2013 Lawrence 2013 Liaw 2011 Onofrei 2004 Wei… Chute 2011 Wilke 2007		Brown 2000 Nadkarni 2014

Appendix B: Electronic Phenotyping studies (cont.)

Statistical Method	Coded Terminology	Text Mining	Hybrid (both)
Specificity 18	Anderson 2016 Asao 2015 Broberg 2015 Coleman 2015 Ferris 2009 Floyd 2016 Fort 2014 Garvin 2013 Lawrence 2013 Liaw 2011 Navaneethan 2011 Onofrei 2004 So 2006 Udris 2001 Wei… Chute 2011 Zhong 2015		Liao 2015 Murff 2011
Recall 6	Lieberman 2003 Thiru 2003 Thiru 2009	Abhyankar 2014 Wei… Chute 2010 Zheng 2014	
Precision 6	Lieberman 2003	Abhyankar 2014 Bromuri 2013 Wei… Chute 2010 Zheng 2014	Agarwal 2016
Accuracy 6	Anderson 2016 Broberg 2015 Lawrence 2013 Liaw 2011 Udris 2001		Agarwal 2016

Appendix B: Electronic Phenotyping studies (cont.)

Statistical Method	Coded Terminology	Text Mining	Hybrid
PPV 29	Anderson 2016 Andrade 2011 Bagheri 2009 Bobo 2011 Borzecki Coleman 2015 Ferris 2009 Floyd 2016 Fort 2014 Garvin 2013 Kleinberg 2013 Lawrence 2013 Liaw 2011 Navaneethan 2011 Onofrei 2004 Rosenman 2013 So 2006 Thiru 2003 Thiru 2009 Udris 2001 Wei… Chute 2011 Wei… Chute 2013 Wei…Denny 2015 Zhong 2015		
NPV 10	Anderson 2016 Floyd 2016 Lawrence 2013 Liaw 2011 Navaneethan 2011 Onofrei 2004 So 2006 Udris 2001		Liao 2015 Nadkarni 2014

Appendix B: Electronic Phenotyping studies (cont.)

Statistical Method	Coded Terminology	Text Mining	Hybrid (both)
AUC/ROC 5	Lawrence 2013 Thiru 2003 Thiru 2009 Wei…Denny 2015		Murff 2011
Simple % match 2	Meyers 2011	Hulse 2013	
Odds Ratio 1			Liao 2015
Coverage 1		Bromuri 2014	
Predicted Prevalence Ratio 1	Asao 2015		
Bayes theorem 1			Abhyankar 2014
Hamming loss/ Ranking loss 1			Bromuri 2014
Total N=61 unique	54	11	5

Appendix C: Value Set Authority Center Downloads

Table 18. *VSAC Value Sets for Diabetes, Chronic Kidney Disease, Stage 5, and Heart Failure*

Diabetes		
OID		
2.16.840.1.113883.3.464.1003.103.11.1001	Diabetes	ICD-9-CM
2.16.840.1.113883.3.464.1003.103.11.1002	Diabetes	ICD-10
2.16.840.1.113883.3.464.1003.103.11.1003	Diabetes	SNOMED
Chronic Kidney Disease, Stage 5		
OID		
2.16.840.1.113883.3.526.2.1035	Chronic Kidney Disease, Stage 5	ICD-9-CM
2.16.840.1.113883.3.526.2.1036	Chronic Kidney Disease, Stage 5	ICD-10-CM
2.16.840.1.113883.3.526.2.1037	Chronic Kidney Disease, Stage 5	SNOMED CT
Heart Failure		
OID		
2.16.840.1.113883.3.526.2.23	Heart Failure	ICD-9-CM
2.16.840.1.113883.3.526.2.24	Heart Failure	ICD-10-CM
2.16.840.1.113883.3.526.2.25	Heart Failure	SNOMED CT

Appendix D: Reference Standard for Diabetes

The Type 2 Diabetes Mellitus phenotype developed by the eMERGE network was used as the reference standard with limited modifications (Pacheco, 2012). This research phenotypes presented some challenges when applied to the context of population health cohorts. It's design as a research phenotype used positive and negative case selection algorithms to find highly specific cases and controls in a sampling methodology. Population health programs require that every patient in an active patient population be included or excluded from a disease cohort. Detailed analysis of patient inclusion and exclusion as it occurred for each rule revealed logic errors, missing or undetected data, contradictory or ambiguous data, and the unavailability of standard LOINC (laboratory codes). Table 19 *eMerge diabetes reference standard findings N= 201,913* details the numbers of patients selected for inclusion based on each of the five phenotype rules. Each row of the table represents a rule. The phenotype algorithm identified 11,278 patients with diabetes in study population of 201,913. Of those, 914 patient had a diagnosis of Type 2 diabetes and had taken Type 2 diabetes medications before starting Type 1 diabetes medications. Another 6,619 patients had a diagnosis of Type 2 diabetes and had taken Type 2 diabetes medications (with no evidence for Type 1 diabetes medications). 1,483 patients had a diagnosis of Type 2 diabetes and abnormal diabetes lab results. 1,129 patients had Type 2 diabetes medication and abnormal diabetes lab, but

no diabetes diagnosis. 1,133 had a two or more Type 2 diabetes diagnosis and Type 1 diabetes medication (no evidence of Type 2 diabetes medication).

Table 19. *eMerge diabetes reference standard findings N= 201,913*

eMerge Phenotype Rule	Count of patients	Type 2 diabetes DX	Type 1 diabetes Med	Type 2 diabetes Med	Type 2 Med prescribed before Type 1 Med	Abnormal diabetes Labs
1	914	Yes	Yes	Yes	Yes	
2	6,619	Yes		Yes		
3	1,483	Yes				Yes
4	1,129			Yes		Yes
5	1,133	Yes	Yes			
Total	11,278					

Logic errors

This algorithm requires that a type 2 diabetes diagnosis or medication is found in the patient record before evaluating the diabetes labs (random glucose, fasting glucose or Hemoglobin A1c). Further analysis of the study data revealed that 1385 patients excluded from the Type 2 diabetes cohort had two A1C results greater than 6.5 at least 90 day apart. The concept of persistence, i.e. abnormal test results that persist over a period of time can be used in phenotyping to reduce the likelihood that a test results was reported in error or is related to a temporary condition that would not benefit from chronic care management.

Therefore, phenotype algorithms that make diagnoses a precondition for evaluating laboratory test may be inadvertently excluding true diabetics. Since diabetic patients receiving effective treatment may have normal glucose and A1c lab results, it may also be overly restrictive for a diabetes phenotype algorithm to require abnormal lab results before searching for a diabetes diagnosis or medications.

Missing or undetected data

In the operationalization of the diabetes reference standard phenotype, the problem list, and ambulatory encounter diagnoses were searched for matching diagnosis codes. There are a number of other locations within the electronic health record where diagnosis data is stored. Discrete diagnosis data is also captured as medical history, inpatient diagnoses, billed diagnoses. Free-text clinical notes have diagnosis information that is largely inaccessible to rule-based algorithms that depend on discrete data.

Contradictory or ambiguous data

Of the 1,385 patients had persistent abnormal A1c (two labs greater than 6.5 over a period greater than 90 days). Further analysis of the data revealed that 715 patients had both Type 1 and Type 2 diabetes definitions. Of those 209 had only a single Type 1 diagnosis and Type 2 diagnosis counts ranging from 1-82.

Challenges with the implementation of LOINC laboratory codes

The eMERGE algorithm specified LOINC terminology codes to identify labs used for diabetic patients. Since lab results in the test EHR did not uniformly capture LOINC codes, local laboratory codes were substituted. This demonstrates that although certified EHR systems must support LOINC standards, the implementation of interfaces and workflows to capture this data may not be implemented at the local level. Table 20 *Local*

diabetes lab result component names used in lieu of LOINC codes details the local names used for each component.

Table 19. *Local diabetes lab result component names used in lieu of LOINC codes*

Laboratory Test	Component Name
Hemoglobin A1C	Glyco HG (A1C)
	HEMOGLOBIN A1C-MEDCOM
	HEMOGLOBIN A1C-QUEST
	HEMOGLOBIN A1C-LABCORP
	HEMOGLOBIN A1C (POCT)
	HEMOGLOBIN A1C / HEMOGLOBIN TOTAL –LABCORP
Glucose	GLUCOSE
	GLUCOSE (POCT)
	GLUCOSE-QUEST
	GLUCOSE-LABCORP
Fasting glucose	GLUCOSE, FASTING -LABCORP
	GTT FASTING-QUEST
	GTT 0-MIN
	GTT 30-MIN
	GTT 1-HOUR
	GTT 2-HOUR

Appendix E: Recommendations for Population Health Programs

Table 21. *Recommendations for population health programs*

	Observation	Recommendation	Goal
1	Current versions of national value sets for diagnosis cohorts were not designed for, and may not be sufficiently accurate to drive clinical interventions. Quality dashboards for clinicians are being developed in parallel with quality dashboards for reporting.	The value set model should be expanded to a phenotype model. Population health programs need to look beyond readily available diagnosis codes to deliver patient cohort definitions designed for actionable interventions.	National phenotypes will increase the efficiency of population health programs by sharing the burden for phenotype development for clinical interventions and quality reporting
2	EHR systems that use terminology vendors to map local codes and proprietary terms to ICD-9, ICD-10 and SNOMED are widely used. VSAC chronic disease value sets that are purportedly synonymous may have significant variation based on the choice of ICD or SNOMED terminology. There was little variation between ICD-9 and ICD-10 value sets in the conditions tested.	ICD should be the preferred diagnosis coding terminology for the retrieval of population health cohorts in national value sets. The hierarchical structure of the SNOMED concepts, coupled with the low interrater reliability for SNOMED coding combine to make the retrieval of SNOMED cohorts more subject to unanticipated variability.	A clinically relevant and consistent set of national value sets and phenotype rules will improve the accuracy and consistency of reported quality measures NOTE: This recommendation would not apply to EHR systems that apply diagnoses in Native SNOMED CT

Table 21. Recommendations for population health programs (cont.)

	Observation	Recommendation	Goal
3	VSAC value sets are published without any quantitative measurement or rating of performance	STARD (Standards for Reporting Diagnostic Accuracy) should inform the essential list of items to report in a study of population health value sets and phenotypes.	Reporting standards for population phenotype studies will facilitate comparisons between different approaches.
4	Although the focus on population health programs is often centered on the technology, the social aspects of system development implementation and measurement apply	Multidisciplinary teams improve the likelihood of success for population health programs.	ROI on population health IT will be increase with the use of good implementation and monitoring processes
5	Use of existing standards such as LOINC and RXNORM are not fully implemented and new standards such as visit types, common definitions of active patients, outcomes such as hemorraghic events are needed.	Quality programs should increase demands for the use of standard terminologies in quality reporting. Where national value sets are not yet mandated, the reporting of value sets used in report cohorts would provide a valuable raw data set with which to inform future phenotype research	Health system interoperability and the foundations of the Learning Health system require the use of standardized terminologies.