Electronic Phenotyping Via the Anchor and Learn Framework with Physical Therapy Emphasis

By

Matt Volansky PT, DPT, MBA

A Dissertation Submitted to

The Department of Health Informatics

Rutgers, The State University of New Jersey

School of Health Professions

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

November 2019

ProQuest Number: 27662811

ProQuest.

ProQuest 27662811

APPROVAL SIGNATURE PAGE

RUTGERS
School of Health Professions

Final Dissertation Defense Approval Form

Electronic Phenotyping Via the Anchor and Learn Framework

with Physical Therapy Emphasis

BY

Matthew T. Volansky

Dissertation Committee:

Shankar Srinivasan PhD

Frederick Coffman PhD

Jane Keehan PT, PhD

Approved by the Dissertation Committee:

_____ Date: 11 / 06 / 2019 _____

_____ Date: 11 / 06 / 2019 _____

_____ Date: 11 / 06 / 2019 _____

_____ Date: _____

_____ Date: _____

ABSTRACT

Background: When delivering evidence-based care at the bedside, learning with anchors is a proven method of efficiently learning statistically driven patient phenotypes. Such libraries currently learn with anchor terms that are universally suspect in their ability to support evidence-based practice for physical therapists (PT) within electronic medical records (EMRs).

Methods: A definition of anchor terms for venous thromboembolism (VTE) was developed using structured and unstructured retrospective data from PT documentation within two separate EMRs. The learned PT specific VTE phenotype anchor terms were compared against the published PT clinical practice guideline and clinician documentation for consistency. The learned PT specific VTE phenotype anchor terms were then evaluated against the published learned anchors derived from physician-based documentation.

Results: Two of the top 25 anchors showed a statistically significant correlation with the presence of VTE: 'vessel' ($P < 0.001$) and 'pe' ($P < 0.05$). The top 20% most frequently appearing learned anchor terms in descending order of total observed frequency was 'boot' (12.2%), 'movement' (10.4%), 'develop' (10.1%), 'cad' (9.5%) and 'pulmonary' (9.2%).

Discussion: This research provides new insight into the relationship between anchor terms and the documentation of PT. The data indicate that the top 20% of discovered physical therapy derived phenotype terms for VTE anchors did not match the existing physician derived phenotype definition for VTE. Based on the existing physician derived anchor terms, clinical decision support tools for VTE would not have been triggered if used by the PT.

Conclusion: The delivery of patient-centered care requires an interdisciplinary team of clinicians to achieve optimal patient outcomes. Evidenced-based practice is enhanced through the presence of clinical decision support tools in the clinical workflow of the modern healthcare system. Before this research, there did not exist an established set of anchor terms with a likelihood of detecting the presence of VTE within the profession of physical therapy. An initial listing of such anchor variables has now been discovered. Further research is needed to expand the ability of machine learning classifiers to identify patients both at risk and with active disease.

ACKNOWEDGEMENTS

I want to acknowledge my heartfelt thanks to the following individuals who have played a significant part in my work.

To my dissertation chair, Dr. Shankar Sirinivasan, for his wisdom and guidance during this period of growth.

To the remainder of my committee, Dr. Frederick Coffman and Dr. Jane Keehan, for their service and significant role in my graduate and professional education.

To my wife, DDr. Kerry Volansky for setting the bar for dedication, love, and lifelong learning.

To my family, Helen, Julie, and Pete, for the unique role that each of them continue to play in my life.

To my family who is no longer with us, Tom, Jim, Bertha, Adam, Helen, and Joel for the special presence which each of them offers me.

To my second family, The Minarczik's and The Homan's for maximizing support and minimizing distractions.

To my professional colleagues who without their support, at one time or another, this work would not have been possible: Dr. Suril Gohel, Dr. Robert Frampton, Dr. Patricia Draves, Dr. Richard Merriman, Cathy Hornbeck, and Desere Hillman.

DEDICATION

To Kerry

Eat, sleep, wake

(nothing but you)

DATA USE AGREEMENT STATEMENT

The limited secondary, retrospective dataset used in this study was provided by Century Oak Care Center part of the Accord Care Community in Middleburg Heights, OH, and is governed by the Data Use Agreement dated January 16, 2019. The purpose of this agreement is to satisfy certain obligations under the Health Insurance Portability and Accountability Act of 1996 and its implementing regulations (45 C.F.R. Parts 160-64) ("HIPAA") to ensure the integrity and confidentiality of Protected Health Information exchanged in the form of a Limited Data Set.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

Chapter I

INTRODUCTION

The government of the United States has become the primary catalyst for advancing

healthcare technology through the establishment of regulatory standards and the

creation of incentives for adoption. With the passing of several significant pieces of

legislation, the role, and importance of health information technology has taken center

stage. See Table 1 for a review of legislation. Most importantly, as a provision of the

American Recovery and Reinvestment Act of 2009, the Office of the National

Coordinator (ONC) was established in law. The ONC provides the U.S. Department of

Health and Human Services with the authority to establish programs to improve health

care quality, safety, and efficiency through the promotion of health IT, including

electronic health records (EHRs) and private and secure electronic health information

exchange. The ONC Health Information Technology Certification Program supports the

Medicare and Medicaid EHR Incentive Programs, which provide financial incentives for

the "meaningful use" of certified EHR technology. These standards look to improve

quality, safety, efficiency and reduce health disparities; engage patients and family;

improve care coordination, and population and public health.[1] The standards provided

for significant "seed" monies, up to $44,000 over five years, for both clinicians and

hospitals to move their clinical operations from paper data into the electronic health record and demonstrate meaningful use.[2] Not surprisingly, hospital adoption of EHRs has increased fivefold since 2008.[3] Government regulation has gained the attention of the private sector for support of these efforts. With 4.5 billion dollars of investment in

TABLE 1. Governmental Infrastructure.	
Governmental Action	**Focus**
The Health Insurance Portability and Accountability Act (HIPAA) of 1996[*]	Privacy and Security of Patient Information.
The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009[†]	Promote the adoption and meaningful use of health information technology.
The Office of the National Coordinator[‡]	Government office created to improve health care quality, safety and efficiency through promotion of health IT, EHRs and secure information exchange.
Section 618 of the Food and Drug Administration Safety and Innovation Act (FDASIA) of 2012 [§]	Development of a risk-based framework for mobile medical applications.
The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA)[‖]	Created the outcomes-based Quality Payment Program for Medicare beneficiary reimbursement based on Advanced Alternative Payment Models (APMs) or The Merit-based Incentive Payment System (MIPS).
The 21st Century Cures Act (Cures Act)[¶]	Improves the flow and exchange of health IT by advancing interoperability, prohibiting information blocking, and enhancing the usability, accessibility, and privacy and security of health IT.

[*] As Accessed on 1/4/19 at: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html
[†] As Accessed on 1/4/19 at: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf
[‡] As Accessed on 1/4/19 at: https://www.healthit.gov/topic/abount-onc
[§] As Accessed on 1/4/19 at: https://www.healthit.gov/sites/default/files/fdasiahealthitreport_final.pdf)
[‖] As Accessed on 1/4/19 at: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/macra-mips-and-apms/macra-mips-and-apms.html
[¶] As Accessed on 1/4/19 at: http://library.ahima.org/doc?oid=302012

digital health startups in the third quarter of 2018, investment in this sector continues with over 50 Billion invested in digital health in the last eight years.[4] With Government initiation for healthcare technology infrastructure adoption and substantial private sector investment, the future for medicine in the United States is most certainly a digital one.

This shift in focus has come at a price. The United States health care system is the costliest in the world, accounting for 17% of the gross domestic product with estimates that percentage will grow to nearly 20% by 2020.[5] In a direct attempt to slow this cost growth The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) amended Title XVIII of the Social Security Act to repeal the Medicare sustainable growth rate (SGR) formula and to strengthen Medicare beneficiary access by incentivizing clinician payments. MACRA encourages clinicians to take part in the Quality Payment Program (QPP) that rewards the creation of value using evidence-based outcomes, instead of the volume of procedures performed. The QPP does this in one of two ways: by a Merit-based incentive payment system (MIPS) or by an Advanced Alternative Payment Model (APMs).[6] The QPPs base their definition of quality on evidence-based measures. This focus on evidence-based practice (EBP) is specifically designed to encourage improvement in clinical practice and require the utilization of advances in technology and interoperability of information exchange.[7]

The change to the reimbursement structure for the treatment of Medicare beneficiaries is pushing both healthcare organizations and providers to focus upon patient-centered care to achieve the Quintuple Aim of Healthcare.[8] The Quintuple Aim refers to the simultaneous achievement of goals for improving the patient experience of care, improving the health of populations, reducing the per capita cost of health care, improving work-life for healthcare staff and equity to prevent health disparities. Without the financial incentives to refocus care on evidence-based outcomes and the quintuple aim created by the QPP, the Medicare insurance program's sustainability and viability were bleak at best.

The shift towards patient-centered care requires an increased level of participation by the patients, as well as the providers of care. Patients experience "value" based on the definition where value is the product of the quality of care plus the patient experience at a given cost.[9] If perceived patient value improves, it can be argued that all players within the healthcare system benefit and the financial sustainability of the system itself increases.[10] In clinical practice, quality is defined as adherence to evidence-based guidelines in the form of Clinical Prediction Rules (CPRs), Clinical Practice Guidelines (CPGs), and Clinical Pathways (CPs). The right treatment, at the right time, for the right patient, is the goal for this type of evidence-based care.[11]

Electronic medical records, although abundant in use, have not begun to incorporate the CPRs and CPGs, which are developing within the field of physical therapy. Physical therapy has moved to graduating professionals with only a clinical doctorate. Beginning with the graduating class of 2020, all of the 242 accredited physical therapy schools in the U.S. will be graduating physical therapists at the doctoral level.[12] The DPT degree is considered a clinical or applied doctorate like those for medicine (MD), dentistry (DDS), education (EdD), clinical psychology (PsyD), optometry (OD), and podiatry (DPM). This push is towards moving the physical therapy profession further into evidence-based clinical practice.[13] Physical therapy evidence-based interventions for falls, balance, stretching, and seating surface assessment occupy 10 of the 20 most accessed reviews in the Cochrane library.[14] Anecdotal reports indicate that the Physical Therapy specialist board examinations have shifted away from expert opinion and towards the incorporation and delivery of clinical care based upon validated CPGs and associated CPRs.[15] If the future standard of general practice is reflected by the current expectations

of the board-certified experts within the field, the profession requires a demonstrated tool that can help close existing clinical performance gaps.

Physical Therapists are expected to function in an expanding role as an independent clinical practitioner who relies on the application of CPGs to obtain accurate diagnoses and outcomes for their patients. Physical therapists are not able to accurately process the many predictor variables needed to use CPGs and CPRs at the point of care while using existing paper-based tools. This makes it unmanageable for clinicians to implement and for the academic institutions to train a new generation of professionals who are ready for advanced clinical practice.

Expert support for the use of CPRs is evident; however, their influence on the educational behavior of new and current physical therapists is limited. In a study of physical therapist's views of CPRs, a survey went out to 292 clinical educators who are responsible for training developing physical therapists who have yet to graduate with their degree.[16] The survey shows that a full 25% had never used a CPR, and furthermore 48% of the respondents had never even heard of a CPR. Twenty-one percent never mentioned them to students, and a further 30% rarely told students about CPRs; only 12% were 'often' encouraging students to use CPRs. The most common reasons for not teaching CPRs were a lack of familiarity and knowledge of CPRs (63%) followed by a desire to encourage students to practice their clinical reasoning rather than using a 'formula' (42%). From the therapists surveyed who had heard of CPRs the most commonly known CPRs were for the identification of injuries to the ankle or foot and the need for an X-ray[17] (Ottawa foot & ankle rules), identification of deep venous thrombosis[18] (Wells Rule) and identification of injuries to the knee and the need for an X-ray[19] (Ottawa knee rules).

5

Awareness of a guideline is necessary for their adoption. Clinicians need to know where to find the guidelines and when to incorporate them into their practice when they are on paper. With electronic CDSS, the appropriate tool is presented to the clinician at the time it is needed, therefore increasing the awareness of the tool over time and the need for it.[20] As demonstrated by a lack of general awareness, support, and training with clinical prediction rules, the specialty of Physical Therapy requires the support offered through the utilization of clinical decision support tools.

Statement of the Problem

At this point, we need to consider the following problem statement:

> **Phenotype libraries currently do not, and in the future, must include expanded heterogeneous definitions to allow for specificity in clinical decision support tool selection and utilization for quality outcomes to achieve optimal patient-centered care.**

There exists a large volume of data that is becoming rapidly available within electronic health records. The exponential growth of big data in healthcare is fueled by governmental planning, oversight, and incentives to both organizations and individual providers. In order to deliver value to the patient, a focus on outcomes must drive quality improvement with a patient-centered care focus.

Clinical decision support services have been slow to work their way into the clinical workflow within the electronic health record despite the identified need for their role in optimal patient outcomes. The focus, instead, has been on process control and not actual outcome management. The available data within the medical record remains inaccessible and locked away as free text inside of siloed proprietary private data farms that are not able to communicate with each other. This lack of data accessibility makes clinical decision rule logic challenging to develop from a computational perspective. Without a universal definition of the specific patient needs identified as phenotypes, which can be used in real-time to fire CDSS at the bedside, it will be impossible to meet the demands for optimal patient-centered care.

Initial attempts at developing CDSS phenotype "triggers" are through the time-consuming process of retrospective, individual chart review using primarily physician

clinical documentation for rule development. Computational methods now exist and are rapidly developing machine learning models for data extraction and analysis.

However, in order to meet the outcome goals of the healthcare system, team-based care and expansion of traditional models of care are necessary. Other primary care professions, such as physical therapy, are developing an evolution of practice which demands revisions to practice acts, educational training, and licensure to allow for the blurring of tasks and responsibility for patient care. Physical therapy, as a profession, has moved to meet this demand by graduating clinicians with clinical doctorates to meet the educational and skill demands to meet this paradigm shift. If real-time CDSS tools are needed by clinicians to achieve optimal patient outcomes, then the patient phenotype "triggers," which call for the CDSS tools must work for all clinicians delivering care.

After performing a comprehensive literature review of clinical decision support tool use and machine learning in physical therapy, this author found no instances of peer-reviewed work in the literature which validates the presence or use of such tools in the profession of physical therapy.

Hypotheses

As previously stated, the phenotype definitions which have been formulated using novel, state of the art machine learning methods, have been too narrow in their focus. Proof of concept for the methods has been performed using only emergency room physician clinical notes as the source for learned classifiers. Therefore, by applying the Anchor and Learn method described by Halpern, Choi, Horng, and Sontag.[21] to include physical therapy clinical note information, the stated phenotype and anchors are thought to demonstrate a significant lack of parity in phenotype definitions. This analysis was focused on the learned anchor for the venous thromboembolism phenotype as an exemplar.

Using the Anchor and Learn methods for phenotype discovery and refinement, the following hypothesis and research objective are offered:

Hypothesis: Using the anchor and learn method, the discovered Physical Therapy derived phenotype definition for venous thromboembolism anchors will not mirror the existing physician-derived phenotypes for venous thromboembolism.

Research Objective: Ultimate adoption of this novel method for use across institutions requires revalidation and expansion by outside researchers. Therefore, research will have the objective of validating these two areas of interest:

Method Validation

1. An investigation will discover if the results are transferable to another electronic medical record dataset within an alternate healthcare facility.

Definition Validation

2. An investigation will discover if the newly acquired anchor observation weights were significantly different when physical therapy notes are utilized

for computable phenotype identification versus those obtained using physician notes.

This research looks to investigate this assertion by looking at a single patient phenotype, Venous thromboembolism (VTE). A comparison was made between the learned phenotype anchors found in emergency physician notes by Halpern et al. and the newly learned phenotype anchors found in physical therapist documentation as a result of this work. Methods for anchor term identification, as described by Halpern et al., were followed to reproduce, as well as, validate these assertions.

The Need for the Study

Next, a case example is used to illustrate for the reader the need for the proposed research. Consider the following case scenario of a patient who presents directly to an outpatient private practice physical therapy clinic, without a physician referral, for an evaluation of their lower extremity and inability to ambulate without pain. This case report, when compared to the existing phenotype definitions derived by physician records, is intended to clearly show the need for an expanded definition of the phenotype to include physical therapy related anchor variables that trigger a CDSS tool for VTE identification and referral.

In this case example, feature weights identified by Halpern et al. using physician notes as a primary source are underlined in **green**. Feature weights using physical therapist notes identified as preparatory to research in this work are underlined in **red**. Feature weights, which overlap between physician and physical therapist documentation, are underlined in **orange**. Table 2 reviews the feature weights using this style.

TABLE 2. Case Example of Feature Weights Physician v. Physical Therapist Source Definitions.							
Physician				**Physical Therapist**			
MD Comments	**Feature Weight**	**Triage Notes**	**Feature Weight**	**PT Comments**	**Feature Weight**	**Therapy Treatment Notes**	**Feature Weight**
DVT	1.43	DVT	1.94	DVT	1.26	fall risk	1.52
lovenox	0.73	leg	0.79	leg	1.89	pain	1.52
filter	0.46	swelling	0.71	ble	1.89	le	1.30
us	0.46	calf	0.64	edema	1.26	transfers max	1.30
anticoagulation	0.45	left Leg	0.56	foot	1.26	edema	1.26
heparin	0.44	rle	0.50	fall risk	0.63	ble	1.08
		lle	0.46			foot	0.65
		clot	0.45				
Abbreviation Key us = Ultrasound rle = Right lower extremity lle = Left lower extremity le = Lower extremity							

Background

The patient is a 44-year-old Caucasian female who presents with a suspected left lower extremity medial gastrocnemius tear. The patient reported an active lifestyle with regular cardiovascular interval and endurance training via stationary and on the road bicycling 4-5 times per week at 100 miles per week for the last ten years.

History of Injury

The patient reported that her initial injury began during a 20-mile bicycle ride. The bike ride consisted of standing sprints and interval training. The patient reported a sudden sensation like "someone kicked her in the back of the left leg." The patient immediately noted the onset of left calf pain, which she rated 5/10 on a visual analog scale (VAS). The next morning the patient reported the inability to weight bear fully on the left lower extremity (LLE) with a reported 8/10 pain on VAS with localization to the left medial head of the gastrocnemius muscle in the LLE. Upon examination, tenderness was noted upon palpation in the entire left medial gastrocnemius muscle, but this tenderness was observed to be exquisitely more painful at the medial musculotendinous junction. A palpable defect was evident at the proximal medial musculotendinous junction. There was no visual discoloration, venous distension, and mild edema was present in the distal left lower extremity above the ankle.

The palpation of the Achilles tendon demonstrated an intact tendon. The peripheral pulses were present and symmetric. Moderate to severe pain was demonstrated with passive ankle dorsiflexion (2 degrees LLE, 18 degrees RLE), as well as, with active resistance to ankle plantar flexion (3-/5 LLE, 5/5 RLE). Circumferential measurement of the left calf was noted to be 3.0 cm larger on the left when compared to the right. The patients'

height was 173 cm; she weighed 54.4 kg. (body mass index, 18.2 kg/m2), temperature 36.7 °C, heart rate 68 bpm, Respirations 16 min, SPO2 99% room air. The Wells Clinical Prediction Rule is used by clinicians to detect the presence of DVT.[22] Currently, the Wells Score was -1. DVT is unlikely. The Wells Score criteria can be reviewed in Table 3.

TABLE 3. Wells Rule Scoring.

Clinical Prediction	Score[†]
Alternative diagnosis to DVT as likely or more likely?	Yes -2
Calf swelling > 3 cm compared to contralateral side?[‡]	Yes +1
Entire leg swollen?	Yes +1
Tenderness along deep venous system?[§]	Yes +1
Pitting edema present?[∣]	Yes +1
Lower Limb recently immobilized? (paralysis, paresis, cast)[¶]	Yes +1
Collateral (nonvaricose) superficial veins visible?	Yes +1
Bedridden recently for > 3 days or major surgery within four weeks?	Yes +1
Cancer? (Active / within 6 months)	Yes +1

Score of 1 or higher – DVT is likely. Refer for further testing.
Score of 0 – DVT is unlikely. Reapply rule if change is observed.

Low = 0 points, Intermediate = 2-3 points, High > 3 points

* Adapted from Wells et al. 1997[22],
† Prediction Score: 0=low, 1-2=intermediate, and ≥ 3= high.
‡ Measured 10 cm distal from the tibial tuberosity.
§ Assessed by firm palpation in the center of posterior calf, the popliteal space, and along the area of the femoral vein distal to the inguinal ligament.
∣ When present bilaterally, is involved > uninvolved?
¶ If walking boot used, consider treating as cast.

The differential diagnosis of calf pain and swelling includes DVT, thrombophlebitis, cellulitis, baker's cyst, muscular injury, tumor or infection, posterior compartment syndrome, arterial aneurysm, and Achilles tendon inflammation or rupture. Several musculoskeletal disorders may present with a similar clinical picture and require careful evaluation to avoid inappropriate investigation and management.[23]

Past Medical History

The patient reported a past medical history of ulcerative colitis (ICD9 556.9, ICD10 K51.90), dysmenorrhea (ICD9 625.3, ICD10 N94.6), migraine (ICD9 346, ICD10 G43.911), and fibroid uterus (ICD9 218.9, ICD10 D25.9). Past surgical history was positive for appendectomy (ICD9 47.0, ICD10 K35.33) in 1987. Past familial history was positive for ischemic heart disease, diabetes, and RLE DVT with resultant venous thromboembolism with the patient's father. The patient reported taking the following

medications: ASACOL EC® 400mg one tab PO bid, AZURETTE® 0.15-0.02mg x21 /0.01 mg x5 one tab PO bid.

Diagnosis

The initial clinical impression yielded a physical therapy diagnosis of calf pain with movement coordination impairment. A suspected grade II muscle strain of the left lower extremity medial head of the gastrocnemius was present. The patient had a primary complaint of LLE gastrocnemius pain and a secondary complaint of gait disturbance with an inability to participate in recreational biking and athletic activity. The patient was not a fall risk. The physical therapy diagnosis was coded as:

- ICD9 844.9 Sprain of knee & leg NEC
- ICD10 S86.112A Strain other muscles(s) and tendon(s) posterior muscle group at lower leg level, left leg, initial.

Treatment

Initial treatment of the medial calf injury included: relative rest, ice, compression, elevation (RICE), and early weight bearing as tolerated during days 3 thru 6 post-injury. Due to ongoing pain and increased fall risk, the patient obtained crutches and initiated partial weight bearing with the use of crutches. The patient ordered a walking boot online to rest the area and wore the brace only during weight bearing. A heel lift was placed inside the walking boot to relieve calf pain. Foot and ankle active range of motion (AROM) was carried out three times a day in a pain-free range. Additional pain management was achieved using ALEVE® (OTC) 220 mg three tab PO initially, then 220 mg one tab PO q6-8hr. Currently, the Wells Score was -1. DVT is unlikely. Reapply rule if a change is observed.

Over the next two weeks, the patients' symptoms slowly resolved. Massage and transcutaneous electrical neuromuscular stimulation (TENS) were added to the treatment plan for residual pain control. The patient discontinued ALEVE® use and resumed full weight bearing without crutches. TheraBand was utilized for active resistance exercise into dorsiflexion, and gradual low-intensity stationary cycling and bilateral heel raises were each performed twice per day.

The patient resumed outdoor biking and spin classes 20 days post-injury, during which time the left calf pain progressively worsened in the medial gastrocnemius region. The patient resumed single crutch use along with ice, elevation, and electrical stimulation for pain management. For the convenience and a faster gait, the patient returned to using a walking boot during community ambulation approximately 8 hours per day on the left lower extremity. Currently, the Wells Score was -1. DVT is unlikely. Reapply rule if a change is observed.

After a week, the calf pain subsided. Eventually, the use of the boot, ice, elevation, and electrical stimulation was discharged. Pitting edema in the left lower ankle and calf was first noted on post-injury day 30. Circumferential measurement of the left calf was recorded to be 3.2 cm larger on the left when compared to the right. There were no reports of warmth, erythema, discoloration, or venous tenderness upon palpation of the left lower extremity at that time. Currently, the Wells Score was -1. DVT is unlikely. Reapply rule if a change is observed.

The patient resumed her typical exercise routine, five weeks post-injury. After attending a spin class approximately one month after the initial injury, difficulties with deep inhalation were noted by the patient, and the patient complained of chest pain. The patient immediately went to the emergency room. The patient was seen in the

emergency department on post-injury day 34 with a primary complaint of shortness of breath.

The patient was diagnosed with left lower extremity proximal and distal deep venous thrombosis with resultant multifocal pulmonary emboli. Transfer to the acute care hospital was made where the patient was on complete bed rest for three days during the initiation of anticoagulation therapy. The patient was discharged home after four days on oral anticoagulants and recommendations for follow up with her pulmonologist.

Discussion

Given this physical therapy case scenario, the clinical feature weight classifiers, which were defined in the original Anchor and Learn method by Halpern et al., show minimal cross over with the proposed clinical feature weight classifiers, which are expected to be present in the physical therapy record. The Wells Score never placed this patient above low risk for DVT. Given that this patient had other comorbidities that would have placed them at risk for DVT, a clinical decision support tool most certainly would have

Focus Area	Term
Diagnosis	Deep Vein Thrombosis (DVT) Pulmonary Embolism (PE) Venous Thromboembolism (VTE)
Physical Therapy	Ambulation Compression Hose Functional Limitation Mobility Motor Activity Movement
Medication	Apixaban (Eliquis) Betrixaban (Bevyxxa) Coumadin (Warfarin) Dabigatran (Pradaxa) Desirudin (Iprivask) Edoxaban (Savaysa) Fondaparinux (Arixtra) Idraparinux Razaxaban Rivaroxaban (Xarelto) Ximelagatran (Exanta, Exarta) YM150 (Darexaban)
Laboratory	International Normalized Ratio (INR) Prothrombin Time (PT)
General Subject Area	Anticoagulant Direct Thrombin Inhibitor Factor Xa Inhibitor

TABLE 4. Estimated Physical Therapy Clinical Features.

From Academy of Acute Care Physical Therapy Clinical Practice Guideline (CPR) for Venous Thromboembolism. (As Accessed on 1/4/19 at https://www.acutept.org/page/VTEGuidelines?)

proven to be a useful adjunctive tool for clinician direct patient monitoring. A phenotypical definition that does not include the identified terms such as fall risk and pain decreases the likelihood of phenotype "triggers" firing appropriate CDSS. This lack of physical therapy terminology inclusion for phenotype "trigger" identification is a genuine issue.

According to the CPR for VTE published by the APTA section on acute care therapy, the following list of terms, as presented in Table 4, can be

TABLE 5. Estimated Physical Therapy Clinical Features Extracted Based on Venous Thromboembolism Clinical Practice Guideline. Part 1	
Focus Area	**Term**
Surgical	Anesthesia time Hip surgery Knee surgery Pelvic surgery Post Op Stroke Surgery Trauma
Physical Therapy	Ankle pumps Calf muscle exercise From acute care From home From long term care Mechanical compression
Mobility	Air travel (2-3 hrs.) Ambulation Bed bound Chair bound Decreased distance (3.1m -10 feet) Lower limb (cast, surgery)
From Academy of Acute Care Physical Therapy Clinical Practice Guideline (CPR) for Venous Thromboembolism. (As Accessed on 1/4/19 at https://www.acutept.org/page/VTEGuidelines?)	

expected to be derived from a physical therapy record when VTE is expected. The physical therapist's responsibility to every patient is 5-fold: Prevention of VTE, Screening for LE DVT, contributing to the health care team in making prudent decisions regarding safe mobility for these patients, patient education, and shared decision making and prevention of long-term consequences for LE DVT. These responsibilities generate additional terms and their derivatives, which are expected to be present in free text and are presented in Table 5 and Table 6 for review.

As can be quickly observed from these compiled tables of expected terms, physical therapist's documentation leans towards the early identification of the signs and symptoms, which are expected to be found in initial stages of thrombosis development. History and physical examination findings will certainly hold most of these free text terms necessary for the identification of the clinical feature weight classifiers. The role of the emergency room physician is to identify the presence of disease or dysfunction, confirm a diagnosis with the aid of diagnostic testing, and to provide subsequent immediate treatment via conservative or direct surgical intervention. Therefore, one would not expect the same

TABLE 6. Estimated Physical Therapy Clinical Features Extracted Based on Venous Thromboembolism Clinical Practice Guideline. Part 2	
Focus Area	**Term**
Medical History	Active cancer or cancer Central venous catheters Fractures History venous thrombosis Inherited thrombophilia Pitting edema Pregnancy or given birth in previous 6 weeks Severe infection Skin induration Venous ectasia Venous ulcer
Physical Therapy	Discoloration (erythema) Homan's Sign Hyperpigmentation Pain on calf compression Passive Dorsiflexion Pretibial edema Prominent superficial veins Redness Villalta scale, score Wells criteria, score
Risk	Age Anesthesia Bed rest Critical care admission Flight travel Hormonal replacement Immobility Oral contraceptives Surgery
Signs & Symptoms	Cramps Falls Heaviness Obesity Pain Paresthesia Swelling Tenderness Unsteadiness Warmth
From Academy of Acute Care Physical Therapy Clinical Practice Guideline (CPR) for Venous Thromboembolism. (As Accessed on 1/4/19 at https://www.acutept.org/page/VTEGuidelines?)	

free text identifiers in the EMR as those found by a Physical Therapist whose job is to identify risk and refer for further follow up.

   We have seen that information is captured in the EMR in a form that is difficult to use in clinical decision support applications for physical therapists, given the current definitions for anchor rules. A significant gap exists when applying the current rules for VTE identification to the physical therapy profession. A complete representation of all data input by all healthcare professionals, including both structured and unstructured data, is necessary to provide patient-centered clinical decision support for optimal outcomes. Halpern et al. have been able to demonstrate a scalable method of building data-driven phenotypes with a small amount of input from a domain expert in the form of anchor variables that can be scaled across institutions. Phenotypes learned in this way are easily scaled to allow use by other institutions. The need for precision medicine to be leveraged and applied by all healthcare professions is a hallmark of the healthcare system of the future. This researcher asserts that the addition of physical therapy data, which was initially excluded from the original dataset defined by Halpern et al., will increase the accuracy and utility of the current anchor definition for the physical therapy profession.

Definitions of Terms
Anchor Observations ("Anchors")[21]

Anchor Observations, also called "Anchors" are characteristics that satisfy two essential conditions to learn a phenotype estimator. The first condition is a high positive predictive value. If an anchor is present, then the patient should almost always have the phenotype. Although anchors must have high positive predictive value, they do not need to have high sensitivity. The second condition is conditional independence. A formal condition that requires that the patient's phenotype is the best predictor of whether the anchor is present in the medical records and that no other data in the document would improve the prediction if the patient's phenotype were already known. Specifying anchors is a manual step because domain expertise is required to identify observations that satisfy the two anchor conditions. After a domain expert specifies the anchors, they are used to build an imperfectly labeled dataset that is passed to a noise-tolerant machine learning algorithm, which learns a more complicated decision rule to estimate the phenotype.

Classifier weight[24]

Classifier weights measure influence. To illustrate how the classifiers change over time, an example of a calculation for a relative influence measure follows: For every patient and every data type, we first compute an unnormalized influence score by taking the sum of weights associated with positive observations in that data type. The influence of a data type on a patient prediction is then computed by taking the absolute values of

the unnormalized influence scores and dividing by the total so that all the values are non-negative and sum to one. The influence of a data type on predictions in an entire patient population is computed as the average influence of the data type on predictions for each of the patients.

Clinical Decision Support Systems[25]

Clinical decision support systems (CDSS or CDS) provide timely information, usually at the point of care, to help inform decisions about a patient's care. Clinical decision support can effectively improve patient outcomes and lead to higher-quality health care. Examples of CDS tools include order sets created for conditions or types of patients, recommendations, and databases that can provide information relevant to patients, reminders for preventive care, and alerts about potentially dangerous situations.

Clinical Practice Guideline[26-28]

"Clinical practice guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances." These guidelines are not fixed protocols that must be followed but are intended for health care professionals and providers to consider. While they identify and describe generally recommended courses of intervention, they are not presented as a substitute for the advice of a physician or other knowledgeable health care professional or provider.

Clinical Prediction Rule[29]

Clinical prediction rules (CPRs) have become popular in the physical therapy literature. CPRs intend to assist clinicians in making a diagnosis, establishing a prognosis, or implementing an intervention. Although it has been suggested that well-constructed CPRs can improve clinical decision making and practice, there is a lack of consensus as to what constitutes a methodologically sound CPR, especially in the derivation stage. In addition to their diagnostic utility, CPRs pertinent to physical therapist practice have recently been developed to assist with subgrouping patients into specific classifications that are useful in guiding management strategies. An advantage of CPRs is that they use the diagnostic properties of sensitivity, specificity, and positive and negative likelihood ratios (LR); thus, their interpretation can be readily applied to individual patients.

CPRs provide practitioners with robust diagnostic information from the history and physical examination that may serve as an accurate decision-making surrogate for more expensive diagnostic tests. For example, the Ottawa Ankle Rules identify only those patients in which the probability of having a fracture is sufficiently large to warrant radiographic imaging, thus reducing costs and avoiding exposing patients to unnecessary radiation. Clinical prediction rules have been developed to improve decision making for many conditions in medical practice, including the diagnosis of

proximal deep vein thrombosis, strep throat, four coronary artery disease, and pulmonary embolism. Clinical prediction rules also have been developed to assist in establishing a prognosis. Examples include determining when to discontinue resuscitative efforts after cardiac arrest in the hospital, determining the likelihood of death within four years for people with coronary artery disease, identifying children who are at risk for developing urinary tract infections, and identifying the characteristics of patients who are likely to develop postoperative nausea and vomiting after anesthesia.

Computable Phenotype[30]

A computable phenotype is a clinical condition, characteristic, or set of clinical features that can be determined solely from the data in electronic health records (EHRs) and ancillary data sources and does not require chart review or interpretation by a clinician. These can also be referred to as *EHR condition definitions*, *EHR-based phenotype definitions*, or simply *phenotypes*. Computable phenotype definitions should be explicit, reproducible, reliable, and valid. Phenotype definitions are composed of data elements and logic expressions (AND, OR, NOT) that can be interpreted and executed by a computer. In other words, the syntax defining a computable phenotype is designed to be understood and executed programmatically without human intervention. Computable phenotype definitions rely on value sets derived from standardized coding systems and may employ hierarchies and weighting factors for data elements.

Deep vein thrombosis (DVT) refers to the formation of one or more blood clots (a blood clot is also known as a "thrombus," while multiple clots are called "thrombi") in one of the body's large veins, most commonly in the lower limbs (e.g., lower leg or calf). The clot(s) can cause partial or complete blocking of circulation in the vein, which in some patients leads to pain, swelling, tenderness, discoloration, or redness of the affected area, and skin that is warm to the touch. However, approximately half of all DVT episodes produce few, if any, symptoms. For some patients, DVT is an "acute" episode (that is, the symptoms go away once the disease is successfully treated), but roughly 30 percent of patients suffer additional symptoms, including leg pain and swelling, recurrent skin breakdown, and painful ulcers. Also, individuals experiencing their first DVT remain at increased risk of subsequent episodes throughout the remainder of their lives.

DVT disease was operationally defined in this research by following the work of McPeek-Hinz, Bastarache, and Denny. Their definition included clots as identified in deep veins, which include internal jugular, super vena cava, inferior vena cava, brachial, radial, ulnar, iliac, femoral, popliteal and profunda femoris veins. The abdominal specific veins splenic, portal, renal and mesenteric were excluded since these are part of the portal circulation. They also excluded these superficial veins including the external jugular, cephalic, basilica, median cubital, small saphenous and greater saphenous. Finally, thrombophlebitis, arterial,

tumor or sinus thrombosis, and manmade venous conduits were excluded from consideration as VTE disease.

Discrimination Statistics[38-39]

The predictive performance of prognostic tests is often reported like diagnostic tests, using estimates of sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve at one follow uptime. These indices of discrimination can be calculated retrospectively and compared when a new prognostic indicator is added to a predictive model, or a prognostic test is compared to predictions made by other methods, including the judgments of experienced clinicians. However, these backward-looking measures of discrimination do not summarize the predicted outcome probabilities and do not directly address questions about the predictions based on a new prognostic test.

Electronic Health Record[40]

Electronic health records (EHRs) are built to go beyond standard clinical data collected in a provider's office and are inclusive of a broader view of a patient's care. EHRs contain information from all the clinicians involved in a patient's care, and all authorized clinicians involved in a patient's care can access the information to provide care to that patient. EHRs also share information with other health care providers, such as laboratories and specialists. EHRs follow patients – to the specialist, the hospital, the nursing home, or even across the country. "The EHR represents the ability to share medical information among stakeholders easily and to have a patient's information follow them through the various

modalities of care engaged by that individual. Stakeholders are composed

of patients/consumers, healthcare providers, employers, and

payers/insurers, including the government."

Electronic Medical Record[41]

Electronic medical records (EMRs) are digital versions of the paper

charts in clinician offices, clinics, and hospitals. EMRs contain notes and

information collected by and for the clinicians in that office, clinic, or

hospital and are mostly used by providers for diagnosis and treatment.

EMRs are more valuable than paper records because they enable

providers to track data over time, identify patients for preventive visits

and screenings, monitor patients, and improve health care quality. "The

EMR is the legal record created in hospitals and ambulatory

environments that is the source of data for the EHR."

Generic Sequence Number[42]

The Generic Sequence Number (GSN), also known as the Clinical

Formulation ID or formerly as GCN Sequence Number, is six digits in

length. The numbers itself do not have significance. First Databank, a

drug compendia publisher, uses a unique GSN to document Drug

attributes, and pricing values are linked to the GSN. The active

ingredient, strength, route, and dosage form are also connected using this

proprietary GSN. The GSN number is the same across manufacturers and

package size. One drug can have multiple GSNs depending upon the

product's available strength, forms, and route of administration. It does

not include indication/class of drug, nor does it allow for ease of

ingredient identification using the code alone. Examples include 25793
Warfarin Sodium (5mg Tablet), 25795 Warfarin Sodium (7.5 mg Tablet),
25790 Warfarin Sodium (10 mg Tablet).

http://reference.pivotrock.net/HealthCareTraining/Drugs/RXC.html

HIPAA "Safe Harbor" Method for Anonymization[43]

1. *Name
2. Address (all geographic subdivisions smaller than state, including street address, city, county, and zip code)
3. *All elements (except years) of dates related to an individual (including birth date, admission date, discharge date, date of death, and exact age if over 89)
4. Telephone numbers
5. Vehicle identifiers and serial numbers, including license plate numbers
6. Fax number
7. Email address
8. *Social Security Number
9. *Medical record number
10. *Health plan beneficiary number
11. Account number
12. Certificate or license number
13. Any vehicle or other devices serial number
14. Web URL
15. Internet Protocol (IP) Address
16. Finger or voice print

17. *Photographic image - Photographic images are not limited to images of the face

18. *Any other characteristic that could uniquely identify the individual – e.g. Free text

= Data field present in data set used in this research and deleted as part of protocol

ICD-10-CM Codes Used[44]

The ICD-10-CM codes used in this research are listed below.

- DVT
    - **I82.409** Acute embolism and thrombosis of unspecified deep veins of unspecified lower extremity
    - **453.40** (ICD-9-CM)

- Embolism
    - **I82.XX** Other venous embolism and thrombosis

- DVT Prophylaxis
    - **Z79.01** Long term (current) use of anticoagulants

- LT use Anticoagulant codes
    - **I82.49** Acute embolism and thrombosis of other specified deep vein of lower extremity
    - **I82.4Y** Acute embolism and thrombosis of unspecified deep veins of proximal lower extremity
    - **I82.4Z** Acute embolism and thrombosis of unspecified deep veins of distal lower extremity
    - **I82.50** Chronic embolism and thrombosis of unspecified deep veins of lower extremity

- D-Dimer Use

    - **R79.1** Abnormal coagulation profile

    - **790.92** ICD-9-CM

ICD-9-CM Codes Used[45]

The ICD-9-CM codes used in this research are listed below. Listing used as suggested by the work of Hinz-McPeek et. al.

- **V12.51** Personal history of venous thrombosis and embolism

    - **453, 453.0, 453.1, 453.2**

- **453.4** Deep vein thrombosis, unspecified

- **453.40** Venous embolism and thrombosis of unspecified deep vessels of lower extremity

- **453.41** Venous embolism and thrombosis of deep vessels of proximal lower extremity

- **453.42** Deep Vein thrombosis, distal

- **453.8** Embolism and thrombosis of other specified veins

    - **453.81, 453.82, 453.83, 453.84, 453.85, 453.86, 453.87, 453.89, 453.6, 453.50, 453.75, 453.51, 453.79, 453.77, 453.52, 453.5, 453.82, 453.7, 453.71, 453.74, 453.76**

- **415.1** Pulmonary embolism and infarction

- **415.11** Iatrogenic pulmonary embolism and infarction

- **415.19** Other pulmonary embolism and infarction

Excluding All

- **452** Portal vein thrombosis

- **451.12** Septic pulmonary emboli

ICD-XX-CM Code[46, 47]

The ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) is a system used by physicians and other healthcare providers to classify and code all diagnoses, symptoms, and procedures recorded in conjunction with hospital care in the United States. This code set went into effect for all services performed / episodes started after 10/01/2015. The "xx" is used as a representative placeholder and can be replaced by "9" or "10" in this work.

Natural Language Processing[48-54]

Natural language processing (NLP) is a collective term referring to the automatic computational processing of human languages. NLP is a branch of artificial intelligence that helps computers understand, interpret, and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

NLP is used to analyze text, allowing machines to understand how a human speaks. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used for text mining, machine translation, and automated question answering.

Bag-of-Words[55-58]

A bag-of-words model (BoW) is a way of extracting features from the text for use in modeling, such as with machine learning algorithms. A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things: a vocabulary of known words and a measure of the presence of known words. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. BoWs use a histogram of the words within the text considering each word count as a feature.

Stanford Part of Speech Tagger[59]

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads the text in many native languages and assigns parts of speech to each word such as nouns, verbs, adjectives, etc. The POS Tagger developed and maintained by Stanford University is a gold standard software tool used for NLP. It is an open-source tool and available at

https://nlp.stanford.edu/software/tagger.shtml

Term Frequency – Inverse Document Frequency (TF-IDF) [60]

The TF-IDF is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The tf-idf weight is composed of two terms: the first computes the

normalized Term Frequency (TF), aka. The number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

NegEx[61]

NegEx algorithm identifies negation in textual medical records. NegEx implements several phrases indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases. It enables word representations in other languages. It is translated to Swedish, French, and German and compared on corpora from each language.

Patient Safety Indicators (PSIs): Agency on Healthcare Research and Quality (AHRQ)[62]

The Agency on Healthcare Research and Quality (AHRQ) has established a set of patient safety indicators (PSIs) to assist in the evidence-based practice for achieving optimal patient-centered outcomes. The PSIs are a set of indicators providing information on potential in-hospital complications and adverse events following surgeries, procedures, and childbirth. The PSIs developed after a comprehensive literature review, analysis of ICD-9-CM codes, review by a clinician panel, implementation of risk adjustment, and empirical analyses. The PSIs

have been adjusted to include ICD-10-CM crosswalks. Patient safety

indicator 12 Perioperative PE or DVT rate is a helpful tool to review

optimized computer definitions for VTE risk.

Patient Value Equation[63]

Value is the

product of

the quality

FIGURE 1. Patient Value Equation.
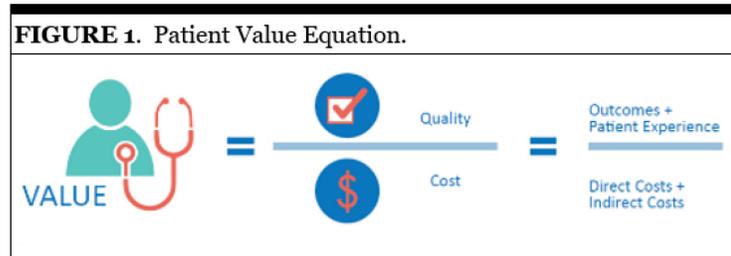
of care, plus the patient experience at a given cost. See Figure 1 for a

graphical representation of the Patient Value Equation.

Primary Care[64]

Primary care is defined as "the provision of integrated, accessible

health care services by clinicians who are accountable for addressing a

large majority of personal health care needs, developing a sustained

partnership with patients, and practicing in the context of family and

community."

Pulmonary Embolism[65, 66]

The most severe complication that can arise from deep venous

thrombosis (DVT) is a pulmonary embolism (PE). PE occurs in over one-

third of DVT patients. A PE occurs when a portion of the blood clot breaks

loose and travels in the bloodstream, first to the heart and then to the

lungs, where it can partially or wholly block a pulmonary artery or one of

its branches. A PE is a severe and life-threatening complication with signs

and symptoms that include shortness of breath, rapid heartbeat, sweating,

and sharp chest pain (especially during deep breathing). Some patients may cough up blood, while others may develop dangerously low blood pressure and pass out. Pulmonary embolism frequently causes sudden death, mainly when one or more of the vessels that supply the lungs with blood are entirely blocked by the clot. Those who survive generally do not have any lasting effects because the body's natural mechanisms tend to resorb (or "lyse") blood clots. However, in some instances, the blood clot in the lung fails to completely dissolve, leading to a severe chronic complication that can cause chronic shortness of breath and heart failure.

Repositories

eMERGE Network[67-70]

The Electronic Medical Records and Genomics (eMERGE) Network is a National Institutes of Health (NIH)-organized and funded consortium of U.S. medical research institutions. eMERGE is a national network that combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine. The Network brings together researchers with a wide range of expertise in genomics, statistics, ethics, informatics, and clinical medicine from leading medical research institutions across the country to research in genomics, including discovery, clinical implementation, and public resources. eMERGE was announced in September 2007 and began its third phase in September 2015.

The primary goal of the eMERGE Network is to develop, disseminate, and apply approaches to research that combine biorepositories with electronic medical record (EMR) systems for genomic discovery and genomic medicine implementation research. Each center participating in the consortium is to study the relationship between genome-wide genetic variation and a common disease/trait. In addition, the consortium includes a focus on ethical issues such as privacy, confidentiality, and interactions with the broader community. In addition, the consortium includes a focus on social and ethical issues such as privacy, confidentiality, and interactions with the broader community. Themes of genomics, bioinformatics, genomic medicine, ethics, data sharing, privacy, and community engagement are of particular relevance to eMERGE. eMERGE current external collaborations include the US Air Force, ENCODE, IGNITE, and the larger ELSI (Ethical, Legal, and Social Issues) community.

SHARPn Project[71-72]

In December of 2010, the Office of the National Coordinator (ONC) announced the Strategic Health IT Advanced Research Projects (SHARP) as part of the federal stimulus project. SHARPn (n is for normalization) is a collaboration of 16 academic and industry partners to develop tools and resources that influence and extend secondary uses of clinical data. The program assembles modular services and agents from existing open-source software to improve the utilization of EHR data for a spectrum of use-cases and focus on three themes: Normalization, Phenotypes, and

Data Quality/Evaluation. The program was assembled into six projects that span one or more of these themes related to research and development. The six projects are (1) Semantic and Syntactic Data Normalization, (2) Natural Language Processing (NLP), (3) Phenotyping Applications, (4) Performance Optimizations, and Scalability, (5) Data Quality Metrics, and (6) Evaluation Frameworks. All of these services are developing open-source deployments as well as commercially supported implementations.

HMORN Network[73, 74]

The HMO Research Network (HMORN) is a member-based network of 17 research centers affiliated with not-for-profit health care systems across the US with the eighteenth site in Israel. These health care organizations all provide comprehensive medical services to enrolled members and patients.

The HMO Research Network (HMORN) Virtual Data Warehouse (VDW) removes duplicative work within a research center by maintaining single extract, transform, and load (ETL) processes for creating commonly used variables in single and multisite research studies. This allows research projects to focus on data development efforts on data not yet included in the VDW. By documenting these pioneering efforts and following VDW documentation guidelines, investigators contribute to expanding VDW coverage. Other projects and other research centers can build on the work of individual projects to expand the VDW data model. Because VDW data files already exist at each site, data query tools may be

used to obtain preparation-to-research data tabulations across multiple sites within days or even hours. Such queries enable investigators to expediently assess the feasibility of research questions and quickly compute statistical power levels.

pCORnet Network[75]

The Patient-Centered Outcomes Research Institute (PCORI) recently launched a new resource known as PCORnet, the National Patient-Centered Clinical Research Network, to increase the speed, efficiency, and relevance of clinical research in the USA. In support of this initiative, PCORI awarded $93.5 million to support 29 health research networks (11 clinical data research networks (CDRNs) and 18 patient-powered research networks (PPRNs)) that together will become a large, interoperable, highly representative, national 'network of networks' for integrating patient-generated data and electronic health information, and conducting comparative effectiveness research (CER).

PCORnet is a novel distributed data network that includes substantial roles and responsibilities for patients and their caregivers in building network capacity, governing and using the health data, and directing a patient-centered research agenda. By encouraging and embedding patients in leadership roles, PCORnet realigns the focus of the existing clinical research enterprise from investigator-driven to patient-centered, thereby advancing the PCORI vision of a paradigm shift that expands the

currently limited roles of the patients and their caregivers in clinical

research participation and decision-making.

Support Vector Machines[76-78]

A Support Vector Machine (SVM) is a machine learning algorithm that

analyzes data for classification and regression analysis that looks at data

and sorts it into one of two categories. An SVM outputs a map of the

sorted data with the margins between the two as far apart as possible.

SVMs are used in text categorization, image classification, handwriting

recognition, and in the sciences.

Triple & Quintuple Aim of Healthcare[79-84]

FIGURE 2. Triple Aim of Healthcare.

Better Health

Patient Centered Care

Value for Money

Best Care

Value and Develop Staff

The *Triple Aim* philosophy of better health, better care and better value for the money keeps patient-centered care as the focus of each encounter
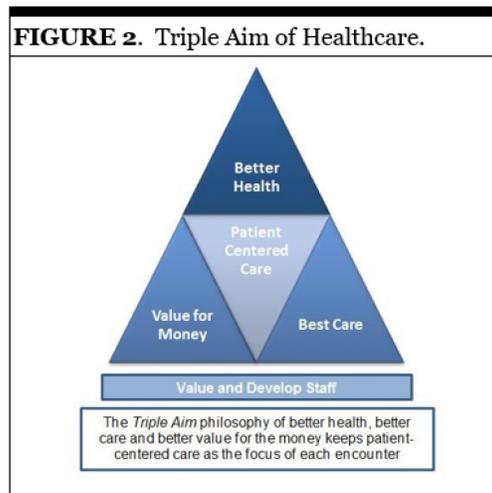
In 2008, researchers at the Institute for Healthcare Improvement (IHI) proposed the "Triple Aim," strategic organizing principles for health care organizations and geographic communities that seek, simultaneously, to improve the individual experience of care and the health of populations and to reduce the per capita costs of care for

populations. See Figure 2 for a graphical representation. Preconditions

for this include the enrollment of an identified population, a commitment

to universality for its members, and the existence of an organization (an

"integrator") that accepts responsibility for all three aims for that

38

population. The integrator's role includes at least five components: a partnership with individuals and families, redesign of primary care, population health management, financial management, and macrosystem integration. In 2010, the Triple Aim became part of the US national strategy for tackling health care issues, especially in the implementation of the Patient Protection and Affordable Care Act (ACA) of 2010.

FIGURE 3. Quintuple Aim of Healthcare.
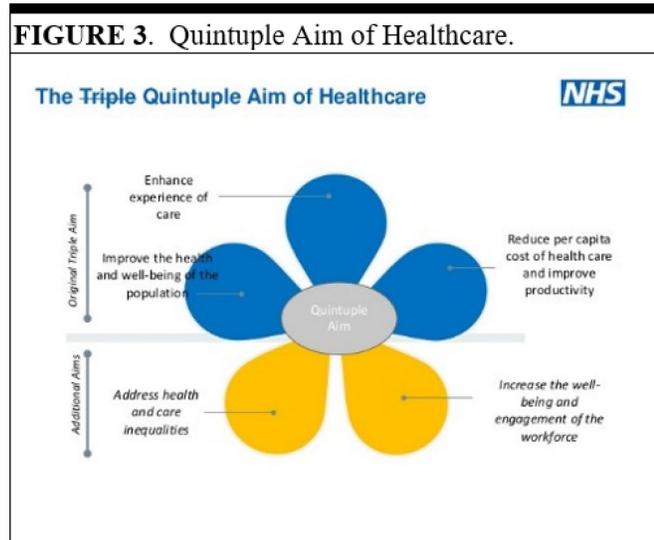
In October of 2016, in a report commissioned by the Agency for Healthcare Research and Quality, Coleman et al. expanded the "Triple Aim" to include two more dimensions: improved work-life for healthcare staff and equity to prevent health disparities. See Figure 3 for a graphic representation of the Quintuple Aim of Healthcare.

Venous Thromboembolism[85]

Deep vein thrombosis (DVT) and pulmonary embolism (PE) are commonly grouped and referred to as venous thromboembolism (VTE).

Wells Rule[18, 22]

The Wells Rule is a well-validated, simple to administer, Level 1 clinical prediction rule. It was developed primarily for use by physicians in ambulatory patients within a suspected first episode of suspected lower extremity deep venous thrombosis. It has been validated in the

TABLE 3. Wells Rule Scoring.	
Clinical Prediction	**Score**†
Alternative diagnosis to DVT as likely or more likely?	Yes -2
Calf swelling > 3 cm compared to contralateral side?‡	Yes +1
Entire leg swollen?	Yes +1
Tenderness along deep venous system?§	Yes +1
Pitting edema present?∣	Yes +1
Lower Limb recently immobilized? (paralysis, paresis, cast)¶	Yes +1
Collateral (nonvaricose) superficial veins visible?	Yes +1
Bedridden recently for > 3 days or major surgery within four weeks?	Yes +1
Cancer? (Active / within 6 months)	Yes +1

Score of 1 or higher – DVT is likely. Refer for further testing.
Score of 0 – DVT is unlikely. Reapply rule if change is observed.

Low = 0 points, Intermediate = 2-3 points, High > 3 points

*Adapted from Wells et al. 1997[22],
† Prediction Score: 0=low, 1-2=intermediate, and ≥ 3= high.
‡ Measured 10 cm distal from the tibial tuberosity.
§ Assessed by firm palpation in the center of posterior calf, the popliteal space, and along the area of the femoral vein distal to the inguinal ligament.
∣ When present bilaterally, is involved > uninvolved?
¶ If walking boot used, consider treating as cast.

emergency room and outpatient physician office settings. Table 3 illustrates the Wells criteria as a series of eight clinical observations and one clinical judgment. Each observation is given a value that is cumulatively added to yield an overall score. The risk for thrombosis is assigned a category of low (0 points), intermediate (2-3 points), or high (>3 points). The associated probability of risk is 3% (95% CI, 1.7%-5.9%), 17% (95% CI, 12%-23%), and 75% (95% CI, 63%-84%) respectively. It is recommended that for patients who score intermediate or high, follow up should be with D-dimer testing and diagnostic ultrasound.

Chapter II

REVIEW OF RELATED LITERATURE

Search Strategies for Literature Review

Materials

The George F. Smith Library of the Health Sciences at Rutgers University was used to search the peer-reviewed literature for instances of physical therapy, electronic medical record, phenotyping, and Anchor Learning methods. It was necessary to limit results that were from the last five years at the time of this search (2013-present). Supplemental searches using the Ohio Link Database of Literature available through the University of Mount Union, (Alliance, OH) and via Google Scholar search engine were also conducted to achieve the desired scope of this review.

Procedure

The following search terms were used in combination: Electronic Health Records OR Medical Record Systems; Computerized OR Artificial Intelligence OR Machine Learning OR Algorithms OR National Language Processing OR Practice Guidelines OR Decision Support Techniques OR Decision Support Systems, Clinical OR Diagnosis, Computer-

Assisted OR Decision Making, Computer-Assisted. This core search was then referenced against the following terms:

Physical Therapy Specialty, Nursing, Physicians, Hospitals. Keywords were mapped to subject and mesh headings. The following databases were searched: Cochrane Library, PubMed, Ovid-Medline, CINAHL. The search included articles published since 2013 and was limited to human subjects, the English language, and available in full text. The reference lists of all primary studies and review articles were hand-searched for additional references and other relevant systematic reviews and are summarized in Figure 4.

FIGURE 4. Literature Search Strategy.

Strategy used for literature review study selection.
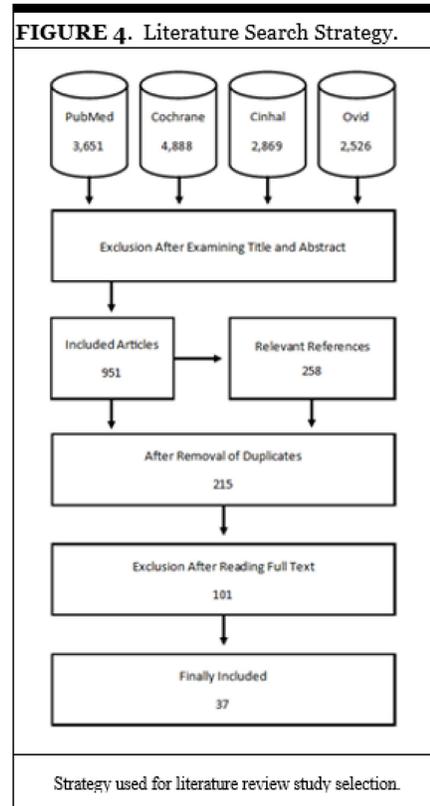
Study Inclusion

Studies were included for review if they met any of the following criteria:

1. The article reviewed a potential best practice of the anchor and learn framework in physical therapy or an alternate discipline.

2. The article reviewed a key term or definition of clinical decision support, phenotyping, clinical practice guidelines, clinical prediction rules, or natural language processing in physical therapy or alternate discipline.

3. There was an analysis of theory as it related to anchor and learn framework, natural language processing, support vector machines, or computer processing of the same.

4. There was a review of possible best practice tools for the creation and deployment of the anchor and learn framework for use by physical therapists in the diagnosis or treatment of patients by another profession.

Literature Review of Selected Value

FIGURE 5. Supportive Literature | Parent – Child Relationship.
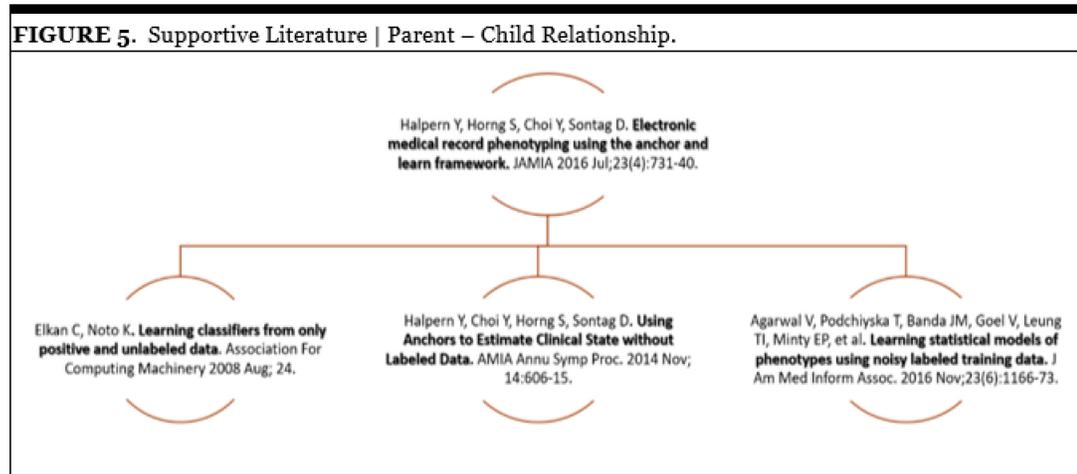
Anchor and Learn

To begin, four key publications work together to form the basis of the theory and methods for the Anchor and Learn method. The scaffolded relationship among the works can be appreciated in Figure 5. These works were reviewed in detail since their methods are being replicated by this researcher.

In the first of four key works, a novel method for gathering data about patients from the electronic medical record (EMR) in a computable phenotype format for use by clinicians at the bedside is described by Halpern, Choi, Horng, and Sontag.[21] The work has a direct application to the ability of active clinical decision support services (CDSS) to guide evidence-based practice (EBP). The authors propose a method of extracting simple facts about patients from the EMR for real-time CDSS specifically for emergency

room physicians. Current phenotype models take time to build since there is a human in the loop who needs to perform the time-consuming task of retrospective chart review. A new method is proposed to allow for free text application with a human still in the loop for only a short period. The newly proposed method continues to benefit from the machine learning predictions of the presence or absence of phenotype grouping of patients. The research focus is on real-time CDSS; however, this anchor and learn method would be useful in retrospective analyses and observational studies. The authors assert that a shift away from retrospective use of International Classification of Diseases (ICD) coding as a primary source when developing phenotypes is needed. Clinical free-text narratives are proposed as a replacement for real-time processing of clinical data for immediate use at the bedside.

The study was performed using the patient data within the EMR of a 55,000 visit per year trauma center and tertiary academic teaching hospital (Beth Israel Deaconess Medical Center, NY). All consecutive emergency department patients for five years between 2008 and 2013 were included in the dataset. Each record represented a single patient visit. No patients were excluded. This led to a total of 273,174 records of emergency department (ED) patient visits. Each phenotype is defined by its anchors, which were derived from eight specific variables: Age, Gender, ICD9 codes, triage vital signs, lab results, medication history, medications dispensed, and free text fields. Binary logic and binning were applied to each of the respective fields. Categorical fields, such as lab values, were binned and converted to ordinal variables. Free text fields used for triage assessment and MD comments were processed using natural language processing using bigram and negation detection before being represented as a binary bag-of-words. A final concatenated feature vector with 21,103 dimensions was used for analysis and

algorithm building. The size of this feature vector is comparable to others reported in the literature. All data analysis was performed using the Python sikit-learn package.

The authors detail the flow of analysis in their methods, which draw from published methods and previous work by the authors.[86, 87, 88] Using these described methods, a path of learning phenotype estimators is described. The critical process steps are to have a single domain expert specify which anchor observations are present in the dataset. Once anchors are identified by the expert, they are used to build an imperfectly labeled dataset, which is then passed to a noise-tolerant machine learning algorithm that learns a complex decision rule to estimate the phenotype. Building each phenotype was reported to take only 10 minutes of physician time using an open-source tool created by the authors.

The authors detail the formation of a phenotype library consisting of 42 clinical phenotypes are presented in Appendix A. Feature weights for each of these phenotypes are reported in Appendix B. An analysis focusing on time-series data results of the predicted patient phenotype is described. The phenotypes chosen were those of immediate relevance in the ED. These features could trigger reminders or CDSS for patient treatment eligibility (anticoagulated, diabetes, history of liver failure), requirements for special monitoring (deep vein thrombosis, suicidal ideation), or the existence of standardized protocols (employee exposure). The authors use the area under the curve (AUC) to demonstrate predictive accuracy during multiple points in time. For all phenotypes, combining free-text and structured data was more informative than either of the two on their own.

In summary, Halpern et al. were able to demonstrate the use of a scalable method of building data-driven phenotypes with a small amount of manual input from domain

45

experts in the form of "anchor variables" that can be shared widely among institutions. The phenotypes can then be implemented as classifiers that can be statistically learned from large amounts of clinical data at each institution. Phenotypes learned this way are shown in this work to be comparable to phenotypes learned with manually identified cases and controls for use in real-time settings and allow for easy scalability.

In the second of the four key works, Halpern, Choi, Horng, et al. set forth the specific rules and definitions for identifying anchors for patient phenotype identification.[88] The utility of this proposed method of predicting patient phenotype is that it can be done by only using a combination of domain expertise and unlabeled data. Other methods rely solely on domain expertise or require manual labeling of positive and negative data examples, which are leveraged by logistic regression, support vector machines, decision trees, and neural networks. Learned classifiers using these methods do not generalize well due to the dependency of the data on which they are trained. Interoperability is difficult using domain expertise or manual labeling. Retraining the learned classifiers at each site requires duplicating the process of labeling data and adjusting the rules. The authors describe in this reported work a methodology for learning to estimate a patient phenotype, consisting of hundreds of clinical state variables gathered from the EMR. This is possible because learned "anchor variables" remain stable between institutions while the rest of the underlying observation model can change.

There are four main contributions of this work. First is the introduction of the concept of "anchor variables." Second is a demonstration of how to use anchors within an unsupervised machine learning algorithm to estimate each clinical state variable without the need for human labeling. The third is a novel user interface developed to help with the ability of a domain expert to choose a good set for anchors for each clinical

state variable. Finally, the last contribution is the evaluation of the created algorithm's performance against nine clinically relevant patient phenotyping tasks.

The "anchor variables" are defined as providing a direct but noisy view of the underlying latent variable, which is to be predicted. Each anchor must satisfy two critical conditions in order to be used to learn a phenotype estimator. The "anchor variables" must have a high positive predictive value; however, they do not need to be sensitive. Once the value of the anchor is known, no other observations provide additional information about the variable. The authors give an example of an anchor phrase "from nursing home" as an example. This phrase, when used in an EMR is highly reliable that the patient lives in a nursing home. However, there are many ways in which "from nursing home" could be written in free text. So, only searching for this single phrase in the EMR would lead to many missed cases. Therefore, a high level of sensitivity for this phrase is not necessary for its use. The crucial second condition is that of Conditional independence. This is described by the authors to mean that "the patient's phenotype is the best predictor of whether or not the anchor is present in the EMR and no other data in the record would improve the prediction if the patient's phenotype were already known."

The authors draw on the work of Elkin and Noto, who treat anchor variables as "noisy" labels and then use them within a learning algorithm.[86] When using a domain expert's knowledge to identify the noisy labels in the data, there is no longer the need to issue a manual label of the data before it can be fed to the machine learning algorithm. This then allows the anchor definition to be very portable between institutions without the need for new labeling work to train the classifiers for a new institution's dataset. The authors describe a computer interface for interactive use by domain experts to specify

anchor variables. "This graphical user interface allows for the specification of the anchor and ease in viewing the learned classifier. Anchors can be specified as words or phrases, and they are interpreted as queries on the free text portions of the medical record. The interface also allows for the incorporation of anchors according to standardized ontologies such as ICD code." The described tool is open source and available on GitHub and can be viewed in Appendix C. Using this tool; the authors report it taking ten minutes for a single reviewer to specify anchor variables.

The methods for learning decision rules with positive anchors are described in detail by the authors. The authors used a collection of 273,174 emergency department (ED) patient records collected from the EMR of a Level 1 trauma center and academic teaching hospital between the years 2008 and 2013. Each record represents a single patient visit, and all consecutive ED patient visits were included in the dataset. No visits were excluded. Data representation and preprocessing show six variables were abstracted from semi-structured sections of the EMR: ICD9 codes from billing information, current medications recorded during medication reconciliation, medications dispensed during the ED course as reported by medication dispensing machines (Pyxis), free text sections formed by a concatenation of chief complaint, triage assessment and physician comments, Age and Sex. Deidentified free text was preprocessed using a modified version of NegEx [61], and negated words were replaced by a new token. A second step of preprocessing collected 1,500 significant bigrams and appended them to the text in order to increase the amount of conditional independence between anchors, which are bigrams and the rest of the text. Medications were represented by a generic sequence number (GSN) and diagnosis by ICD9 codes. Age was discretized by decade with a binary indicator for each decade. Patients are represented

as a binary feature vector representing the presence or absence of each distinct diagnosis code, current medication, dispensed medication, word, discretized age value, and sex. Observations that occurred in fewer than 50 patients in the entire dataset were discarded. A final feature vector size of 20,334 was reported.

The authors go on to discuss anchor tool used by a single emergency physician specified anchors for each clinical state variable using the author's custom anchor elicitation tool with access to a database of 20,000 unlabeled patients chosen at random from the full patient set. Each anchor-based predictive model was only built once. Using the interactive tool, the total time to specify anchors for all nine models was approximately 5 hours.

A machine learning (ML) comparison was then made between (1) the classifiers learned using anchors to (1) a simple rule-based baseline and (2) a supervised machine learning baseline that uses a subset of the collected gold standard labels for training. The evaluation was reported on nine separate estimation tasks, one for each clinical state variables (didFall, hasCardiacEtiology, hasInfection, fromNursingHome, hasCancer, hasPneumonia, isAnticoagulated, is Immunosuppressed, hasSepticShock). (1) The rule-based baseline predicts positively when at least one anchor is present and negatively otherwise. This approach requires no training and is evaluated on the entire labeled set. (2) Evaluation of the supervised baseline is reported using 4-fold cross-validation. In each experiment, the labeled patients are divided into four equal-sized test sets. For each test set, a classifier is trained using a portion of the 75% of patients who are not in the test set ("training patients") and then used to predict for the 25% of patients designated as a test. The results are averaged across the four test sets, giving an estimate of the performance on the entire labeled dataset. Each classifier of the supervised

baseline is learned using at most 3,000 training patients. The supervised baseline was learned with logistic regression using the scikit-learn package in Python.[89] The authors used 5-fold cross-validation within the train set to choose parameters, trying all combinations of the regularization constant (options are $\{10^{-6}, 10^{-5},...,10^{6}\}$) and the norm used in regularization. Choices for norm are L1 or L2. L1 encourages the learned classifier to use a minimal number of features by penalizing the sum of absolute values of the regression weights. L2 avoids overly emphasizing any one feature by penalizing the sum of squares of the regression weights. The authors reduce the regularization of the bias parameter by setting the "intercept_scaling" parameter to 1,000. The Anchor method is trained using the specified anchors, and 200,000 examples chosen randomly from the unlabeled dataset and tested on the entire labeled dataset. Scikit-learn was used to fit logistic regression models as in the supervised setting but holding the regularization norm fixed as L2 and doing cross-validation over the regularization parameter. Since the logistic regression models learned in the anchor method are meant to predict the presence or absence of the anchor, the cross-validation technique to chose parameters also uses the presence or absence of the anchor to measure performance, requiring no ground truth labels. Performance is measured using the area under the ROC curve (AUC), a measure of overall quality of a ranking predictor. Estimating constant "C" in step 2 of the anchor algorithm is not necessary to obtain a ranking, so this step was omitted. In the rule-based approach, ties are broken by counting the number of distinct anchors present in the patient record. In the anchor approach, ties among patients with anchors are broken according to the predicted probability of the latent variable ignoring the presence of the anchors.

The results demonstrate that across the nine clinical state variables, the anchor-based unsupervised learning algorithm obtains prediction accuracy comparable to and in many cases, better than a supervised prediction algorithm.

In the third of four key works, Agarwal, Podchiyska, Banda, et al. describe the need for clinical phenotype descriptions which work across clinical data warehouses.[87] The authors assert that "the rate-limiting step in the compilation of cohorts for clinical research is the generation of clinical phenotype descriptions," and that manual creation of training sets for machine learning is time-intensive. The concept of learning phenotypes with "noisy" labels, which are learned rather than selected by content experts as a gold standard, is explored within this work. The authors demonstrate how such phenotypes can be learned when using rule-based definitions published by the Electronic Medical Records and Genomics (eMERGE) database and the Observational Medical Outcomes Partnership initiatives (OMOP). These rule-based definitions require cross-validation via manual chart review by other partner institutions for phenotype selection to be interoperable. The authors use the Type 2 diabetes Mellitus definition from the eMERGE database, and the MI definition from the OMOP database as examples of acute and chronic definitions for comparison with the new method of phenotype identification via noisy labeling.

A unique feature of this work was that it revealed the computer hardware used by the researchers. This hardware was reported to have had 16 cores and 170 GB of RAM. Although not complete in its detail, this reporting of hardware specification gives the reader the ability to ascertain what hardware will allow for feature selection via machine learning. The authors also gave the time range of 2 to 3 hours to complete each phenotype model to be trained. These are methods that are rarely reported in the

literature and are unique here. They can be used as a benchmark for future researchers. The work also helped to document the number of feature dimensions that will need to be utilized for adequate machine learning. The size of features obtained was concatenated to be 23,717 dimensions for MI and 25,045 dimensions for Type II Diabetes Mellitus. The effort for cross-referencing via these two established database warehousing models necessitated intensive clinician input for keyword list validation for phenotype identification.

This work, although essential to show that noisy labeling can be as accurate as a human-initiated gold standard chart review, is not practical. The time spent along with the method in which the phenotype definitions needed to be cross-referenced to these standardized databases is too burdensome. Institutions looking to replicate this work need to have a working knowledge of the data warehouses used and have both a dedicated technical and clinical team available to validate the procedures. Ultimately, the authors assert, "The assumption is that if a patient exhibits the phenotype of interest, then a doctor is likely to mention it in their notes and that if a highly specific phrase is found, the patient is likely to have the phenotype." The simplicity in the logic of this statement as a summary of this work allows it to have a broad application going forward.

Finally, in the crucial fourth work, Elkan and Noto focus on algorithms that learn binary classifiers, which typically consist of two sets of examples.[86] One set that contains positive examples of the concept to be learned and the other a set of negative examples. It is often the case that training data for the algorithm do not fit this definition. Training datasets are quite often a combination of an incomplete set of positive examples and a set of unlabeled examples containing both positive and negative examples. In this work,

Elkan and Noto identify how to solve this problem. This work is used as a central component of the Anchor and Learn method described by Halpern.

Most research on training classifiers in data mining and machine learning assumes that clearly defined cases of negative examples are available. Databases exist that detail positive examples of items. It would be useful to learn a classifier that can identify additional items that should be included in this list of positive examples. The goal of the work by Elkin and Noto was to discover a method for answering the question, which is not possible to be answered by human experts.

In this work, positive examples need to be "completely selected at random." If positive training examples are labeled at random, then "conditional probabilities produced by a model trained on the labeled and unlabeled examples differ by only a constant factor

FIGURE 6. Two Classifiers Used to Produce a Two-Dimensional Gaussian.

Data points lie in two dimensions. Blue pluses are positive examples. Red circles are negative examples. The large ellipse is the result of logistic regression trained on all the data. It shows the set of points x for which $p(y|x) = 0.5$ is estimated. The small ellipse is the result of logistic regression trained on positive labeled data versus all other data, then transformed following Lemma 1. The two ellipses represent similar classifiers in practice since they agree closely everywhere that data points of both classes have high density.

from the conditional probabilities produced by a model trained on fully labeled positive and negative examples." This is similar to using a nonlinear kernel with support vector
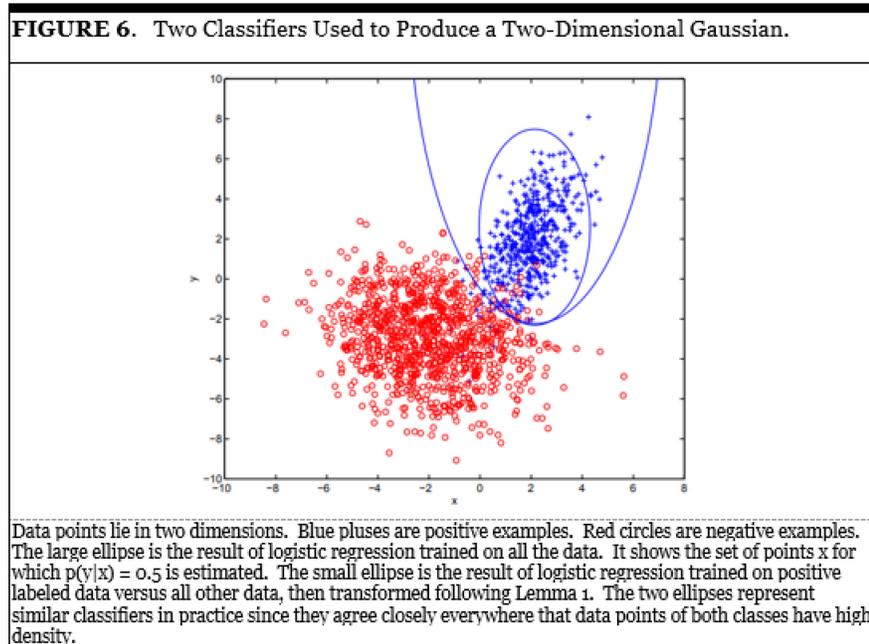
machine. Using this technique is also faster because correct weighting factors are computed directly.

The authors provide a mathematical proof of their work which they call Lemma 1. They offer proof of their example by training two classifiers to produce a two-dimensional Gaussian, as reviewed in Figure 6. If you take a training data set of 1,500 examples (500 positives and 1,000 negatives) and train classifiers on the full dataset (500 positive Vs. 1,000 negative), you get a conditional probability of 0.5. Each point on the curve has a predicted probability of 0.5 of belonging to the positive class. If you take the same training data set of 1,500 examples and only use a "purely random sample" from 20% of the positively labeled data and use that for training classifiers (100 positives Vs. 1,000 negatives + 400 positives). The actual probability of being correct is 20% (0.02). The estimated probability using the author's proposed method using only positive and unlabeled data was found to be 19.28% (0.1928). This is very close to the actual value, which was 20% or (0.02). The authors go on to assert that, "despite the two ellipses derived are visually different, they correspond closely in the area where both positive and negative data points have high density" so they represent similar classifiers."

The authors go on to detail an application of Lemma 1 to real-world data. They describe a detailed proof of learning from positive and unlabeled data in document classification. Their experiment uses an annotated protein sequence database maintained by the Swiss Institute for Bioinformatics and the European Bioinformatics Institute called the Swiss-Protein Knowledgebase (SWISS-PROT). They used a training set in each trial consisting of 90% of the data, with 70% used for training and 20% for validation in each trial using an SVM with soft margins and a linear kernel. The authors

compare four different approaches and compare the accuracy of each of the four classification methods used. For each method, the receiver operating characteristic (ROC) curve was plotted, F1 score or error rate for natural thresholds for yes/no classification, and recall at a fixed false-positive rate given by a human expert in the application domain. The authors go on to give a simplified example to explain the resultant predictions of the four approaches. "Suppose that a human expert will tolerate 100 negative records. Then the expert will miss 9.6% of positive records using the three biased SVM methods, but only 7.6% using the reweighting method, which is a 21% reduction in error rate, and a difference of 55 positive records in this case."

This work is essential because in medicine when we are looking for positive examples of a patient who have a specific condition, we may always find additional factors which would include patients based on comorbidity or other such attributes in their current or past histories which would add to the probability of expression of the condition. Therefore, we always have positive examples, but there may be other expressions not identified in the negative dataset, which are unlabeled or unidentified. Therefore, being able to learn classifiers using positive examples and unlabeled datasets is extremely important.

The scaffolded thought used by Halpern et al. when formulating the Anchor and Learn methodology relies heavily upon the works of Noto and Elkin and Agarwal et al. The methodologies used in these two published works allow for the Anchor and Learn method to learn phenotypical information from free text within the electronic medical record with only minimal input from expert reviewers. It also allows for the replication of the work by alternate researchers using any electronic medical record if the phenotype definitions are known. It is for this reason that these four works were reviewed in detail.

Systematic Review

There was a single systematic review available for analysis. Shivade, Raghavan, Fosler-Lussier, et al. searched the full text literature in every article published in the three years 2010-2012 in four major biomedical informatics journals: (1) Journal of American Medical Informatics Association, (2) Journal of Biomedical Informatics, (3) Proceedings of the Annual American Medical Informatics Association Symposium, and (4) Proceedings of Clinical Research Informatics Conference for studies describing systems or reporting techniques for identifying cohorts of patients with specific phenotypes.[90] Rule-based systems for automatic identification of patients with a typical phenotype are on the decline. Statistical analyses, machine learning, and natural language processing techniques are on the rise. The authors recommend that standardization of data using common terminologies is a critical step towards portable interinstitutional solutions. Tools that are commonly accepted to define phenotypes are lacking. The authors strongly suggest that: "the biomedical informatics community should firstly channel efforts towards making open-source tools that are well documented, maintained and easily available to users. Secondly, there should be a focus on reporting the performance of these available tools before a new one is developed." The authors, after reviewing the literature, also feel that "there should be a focus on developing systems that make holistic use of the electronic health record in characterizing a patient for phenotyping purposes." The use of additional data sources is necessary to identify patients for many phenotype use cases: Demographics, Medications, Lab reports, Vitals, Clinical, Diagnosis, Treatment, Notes, Genomic, Other (primary recommended data sources for electronic health record use in phenotyping). The authors also assert the recommendation that "if generalizable solutions are to be

developed, the use of statistical and machine learning methods are necessary. Such efforts to port phenotyping algorithms across multiple sites should be expanded to test their robustness and scalability."

Phenotyping

Concept

The concept of interpretability is discussed by Gehrman, Dernoncourt, Li et al. in their work comparing deep learning and concept-based methods for patient phenotyping from clinical narratives.[91] Interpretability is how easy one can understand how a model arrived at a prediction.[92] If it is not clear how a prediction model was constructed and inherent bias in this coding is not illuminated, then this is a possible drawback from adoption and usability by clinicians and end-users. This concept is gaining popularity. The European Union is considering regulations to require algorithms to be interpretable.[93] The authors examined ten phenotypes using 1,610 discharge summaries from both nurses and physicians who were admitted to the adult intensive care unit at the Beth Israel Deaconess Medical center from 2001 to 2012. Using a phrase dictionary, the presence of ten individual phenotypes were annotated in each discharge note using a standardized definition for inclusion and are summarized in Table 7. Each note was manually labeled twice by seven researchers (2 clinical researchers, two junior medical residents, two senior medical residents, and one intensive care medicine physician). The authors then compared this manual assignment to the following methods: convolutional neural networks (CNN), Bag of words (BoW) + logistic regression, n-gram + logistic regression, the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) using full output, and cTAKES using filters. The authors support the use of convolutional neural networks (CNNs) as a superior approach to text-based phenotyping

over extraction-based and n-gram based methods. This work illustrates the need for moving away from manual extraction of phenotype phrases and towards computer-assisted classification and prediction. This can be seen by the authors reporting it taking ten minutes per chart

TABLE 7. Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives. [91]

Phenotype	#Positive	Inter-rater Agreement	Definition
Adv. / Metastatic Cancer	161	0.83	Cancers with very high or imminent mortality (pancreas, esophagus, stomach, cholangiocarcinoma, brain); mention of distant or multi-organ metastasis, where palliative care would be considered (prognosis < 6 months).
Adv. Heart Disease	275	0.82	Any consideration for needing a heart transplant; description of severe aortic stenosis (aortic valve area < 1.0cm2), severe cardiomyopathy, Left Ventricular Ejection Fraction (LVEF) <= 30%. Not sufficient to have a medical history of congestive heart failure (CHF) or myocardial infarction (MI) with stent or coronary artery bypass graft (CABG) as these are too common.
Adv. Lung Disease	167	0.81	Severe chronic obstructive pulmonary disease (COPD) defined as Gold Stage III-IV, or with a forced expiratory volume during first breath (FEV1) < 50% of normal, or forced vital capacity (FVC) < 70%, or severe interstitial lung disease (ILD), or Idiopathic pulmonary fibrosis (IPF).
Chronic Neurologic Dystrophies	368	0.71	Any chronic central nervous system (CNS) or spinal cord diseases, included/not limited to: Multiple sclerosis (MS), amyotrophic lateral sclerosis (ALS), myasthenia gravis, Parkinson's Disease, epilepsy, history of stroke/cerebrovascular accident (CVA) with residual deficits, and various neuromuscular diseases/dystrophies.
Chronic Pain	321	0.83	Any etiology of chronic pain, including fibromyalgia, requiring long-term opioid/narcotic analgesic medication to control.
Alcohol Abuse	196	0.86	Current/recent alcohol abuse history; still an active problem at time of admission (may or may not be the cause of it).
Substance Abuse	196	0.86	Include any intravenous drug abuse (IVDU), accidental overdose of psychoactive or narcotic medications,(prescribed or not). Admitting to marijuana use in history is not sufficient.
Obesity	155	0.94	Clinical obesity. BMI > 30. Previous history of or being considered for gastric bypass. Insufficient to have abdominal obesity mentioned in physical exam.
Psychiatric disorders	126	0.91	All psychiatric disorders in DSM-5 classification, including schizophrenia, bipolar and anxiety disorders, other than depression.
Depression	460	0.95	Diagnosis of depression; prescription of anti-depressant medication; or any description of intentional drug overdose, suicide or self-harm attempts.

The above is a representation of manual chart review of paper discharge summaries by 7 researchers representing over 500 person hours to establish phenotype inclusion of patients within 1,600 health records during the period of 2001-2012.

* Adapted from Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction-based methods for patient phenotyping from clinical narratives. PLoS ONE. 2018;13(2):e0192360.

to extract the ten phenotype terms which account for over 500 hours to code the 1,610 discharge summaries once. Recalling that this practice took two rounds of review, using an extraction-based method for phenotype identification accounts for over 1,000 person-hours to complete. The authors agree that the accuracy of patient phenotyping can be improved without the need for a phrase-dictionary as input.

Kirby, Speltz, Rasmussen, et al. created the Phenotype Knowledgebase (PheKB) as a workflow management system and learning center supporting the creation, validation,

and dissemination of computable algorithms between member institutions.[94] It can be accessed at https://phekb.org/. Within the PheKB database, the most common data modality used for the creation of algorithms by members of the knowledgebase were ICD-9 codes, medication data, NLP, CPT codes and laboratory test results. The PheKB has over 250 member institutions. Each member institution must possess the resources for continual data governance, enterprise-level network server/Ruby web service administration and end-user support, metadata encoding capacity and versioning of uploaded algorithms. Given these requirements, the barrier for entry is high, with this model of autonomously developing and sharing phenotype algorithms efficiently. The authors report that the output is variable among member institutions. Due to differences in electronic health record architecture, local customization and workarounds are required. This causes some algorithms within the PheKB to perform with varying levels of accuracy from one site to another. PheKBs primary utility seems to be the ability to allow for a centralized knowledge base to share methods, best practices, and workflow with other member institutions. It demonstrates a significant step forward to allow for the centralized curation of data algorithms. Given the proprietary nature of the technologies used to create, model and maintain these definitions, the barrier to entry is very high within the cash strapped healthcare system.

Electronic Health Record

CONSEQUENCES

Chase, Mitrani, Lu, et al. reported that diagnostic accuracy might be improved by mining patient's clinical notes in the electronic medical record (EMR) for signs and symptoms for specific diseases using natural language processing (NLP) methods and machine learning (ML) which could shorten time to diagnosis.[95] This new study used

internet search histories of patients with pancreatic cancer to identify standard search

terms used to identify the presence of pancreatic disease before patients knew that they

had the disease.[96] The authors suggest that low prevalence conditions might be the best

targets for this type of novel alert strategy. For example, Multiple Sclerosis prevalence is

only 0.8% making it an outlier diagnosis in early stages. Having the ability to predict the

presence of a disease state early may lead to optimal patient outcomes, decreased

morbidity and prolonged quality of life.

   Cabitza, Rasolini, and Gensini presented their viewpoint of unintended consequences

of machine learning in medicine.[97] The authors warn of the potential over-reliance on

machine-based predictive tools in medicine. Their worry is that of the "deskilling" of the

healthcare workforce over time. They report that the reduction of the level of skill

required to make a diagnosis or select the appropriate therapy once such tasks are

automated poses a risk when the technologies do not function correctly or breakdown.

The authors describe studies in which there was a 14% decrease in diagnostic sensitivity

in reading computer-aided mammograms and a 9% decrease in reading computer-aided

electrocardiograms by medical residents.[98, 99] The potential is to focus solely on the data

and not on the many medical nuances which are very difficult to describe in computable

language. The absence of such information may lead to making incorrect interpretations

due to a lack of context. The authors cite an example of this where the use of a machine

learning prognostic algorithm yielded correct results with a counterintuitive and

dangerous output based on bias in the data.[100] In a hospital system, an unaccounted-for

operational workflow caused the mortality risk prediction to make decisions about

whether to provide treatment on an inpatient or outpatient basis for patients with

pneumonia. The described hospital developed a best practice care pathway that sent

patients with a history of asthma who presented with pneumonia directly to the intensive care unit to prevent complications. Once the machine learning algorithm was trained on this data, it correctly learned to predict that patients diagnosed with both asthma and pneumonia had better outcomes than those diagnosed with pneumonia without a history of asthma. This blind spot in the data led to the inability to identify that this best practice would lead to the algorithm "correctly misinterpreting" the presence of asthma as a protective variable. The authors strongly encourage the transparency of the data used to train and make machine learning predictions within healthcare settings.

HUMAN REVIEW

Estaban, Tablado, Ricci, et al. showed that manual chart review was used as the gold standard against the performance of a rule-based algorithm to identify clinically relevant cardiovascular and cerebrovascular disease cases using data from the electronic medical record (EMR).[101] Three family physicians selected clinically relevant cardiovascular and cerebrovascular disease terms from the international classification of primary care, ICD-10-CM and SNOMED-CT sources. A term list that identified each of the disease processes, signs, symptoms, diagnosis or procedures was curated and agreed upon by the expert reviewers. An algorithm was created using this data. The performance of the algorithm to predict the presence or absence of the two disease states was compared to expert review. The electronic medical record of 1,106 patients was reviewed. No mention was made of time spent curating the manually reviewed gold standard cases used. However, using ten minutes per chart for extraction, as reported by Gehrman, Dernoncourt, Li et al., an estimated amount of time could reasonably be assumed to be 184 hours per reviewer. This would be over 500 hours of focused human review when combined. The algorithm showed high sensitivity (0.99, 95% CI 0.938-0.9971) and

acceptable specificity (0.86, 95% CI 0.818-0.895) for detecting cases of cardiovascular

disease and cerebrovascular disease combined. This work demonstrated that rule-based

algorithms are valid, but confirms that they are very time-consuming.

Kukhareva, Staes, Noonan, et al. proposed that when chart reviews are done by a

single-reviewer with phenotyping support before review, it can be done quickly,

accurately, and at a lower cost than without electronic phenotyping support.[102] This

work demonstrates utility for institutions where human resource support for such work

is underfunded or limited. Such methods proved to be of most benefit for the accurate

identification of rare exclusion data within the electronic medical record. Recorded

accuracy of unassisted manual chart review was 92.47% and 98.3% when using the

electronic phenotyping control review strategy. Interestingly, the single reviewer in this

study did not have formal medical education and had 16 years of experience in

conducting validations for quality improvement/assessment, billing, and coding. Time

to review the chart for phenotypic information was reported to be, on average, 76.78

minutes on a dataset of 3,104 cases. Total time spent gathering gold standard chart

reviews assessing the accuracy of electronic phenotyping results by a single analyst was

significant at 3,972 hours. The authors also calculated the cost of human versus

phenotyping support in conjunction with human review ($48.54 versus $63.56, p = 0.16)

In this case the compliance specialist performed all reviews as part of their established

job duties reviewing for National Committee for Quality Assurance's Healthcare

Effectiveness Data and Information Set (HEDIS) quality measures. There is one

identified downside to using this approach proposed by the authors. The human

reviewer tended to agree with the electronic phenotyping initially offered despite those

results being incorrect. (56.67% vs 80%, N = 55 p = 0.07). This finding is another

example of the potential for "deskilling" of the healthcare workforce, as warned in work by Cabitza, Rasolini, and Gensini.

Liao, Cai, Savova, et al. were looking to design a phenotype classification algorithm for rheumatoid arthritis that would identify patients with a high positive predictive value >90% for the disorder.[103] As part of the Informatics for Integrating Biology and the Bedside (i2b2) project, the authors applied a single general approach to developing several phenotype algorithms for depression, diabetes mellitus, inflammatory bowel disease (Crohn's disease and ulcerative colitis), multiple sclerosis and rheumatoid arthritis. The authors describe an extensive multidisciplinary team that is required to complete this work. Team members included clinical investigators, biostatisticians, EMR informaticians, and natural language processing experts. The authors document a basic "flow" of methods that can be used by future researchers. Their findings demonstrate that the incorporation of natural language processing (NLP) improved the performance of all the algorithms. The combination of both structured and NLP data performed the best across all algorithms used. NLP improved all algorithms using structured data by increasing sensitivity while maintaining or improving accuracy. This is because NLP added independent predictive variables to the algorithm. The authors make recommendations for the use of relational databases for the storage of structured data from within the EMR. The authors also discuss how the positive predictive value of an algorithm depends on the prevalence of the disease. They suggest that as a first development step for any phenotype algorithms, one should identify positive cases and negative cases from within the database. They are also quick to identify that a significant limitation of work using this method is the time and resources needed to identify and extract the variables for the algorithms. The authors state that any institution with an

EMR can develop phenotype algorithms. However, tapping into these data requires many additional technical team members to transform the data into a useable structure and to reformat, manage and extract the data. Another major limiting factor was attempting to map terms to NLP concepts. This work is an example of a successful, albeit time and labor-intensive method for extracting EMR data in order to gain phenotype information on an identified cohort.

Alnazzawi, Thompson, Batista-Navarro, et al. present a new expansion of text mining to develop a new corpus called the "PhenoCHF." [https://code.google.com/p/phenochf-corpus.](https://code.google.com/p/phenochf-corpus) The novelty of this work is that it integrates text from both discharge summaries in the electronic medical record (EMR) and scientific articles from the literature.[104] The discharge summaries contained many challenges for phenotype associations such as unstructured grammatical sequences, domain-specific abbreviations, complex sentences, and spelling errors. The literature-based free text was void of a majority of these errors. The literature corpus used consisted of the most recent ten full-text articles retrieved from the PubMed Central Open Access database at the time of corpus collection. Phenotype disease associations were done by two physicians before training machine learning algorithms for identification. For machine learning testing, each part of the corpus was divided into a training set (80%) of the data and a test set (20%) of the data. Using text outside of the EMR to facilitate the expansion of search terms has shown that rule-based systems perform best. Machine learning methods are comparable but need both databases to achieve good results. This article is essential because it demonstrates the ability to extract comprehensive phenotype information from multiple sources with differing characteristics. A system that is trained to recognize phenotype information in

an EMR can achieve good performance levels (F scores from 0.87 - 0.92) when the same task is used on literature articles.

<div align="center">NATURAL LANGUAGE PROCESSING</div>

Travers and Haas began the work of using NLP for building nursing vocabulary indexes within the emergency department (ED) chief complaint (CC) note sections of the electronic health record.[105] The CC drives patient flow during the triage process in the ED. The authors looked to map the CC, which is in free text form, to existing terminology databases. The authors mapped the identified ED CC concepts using the Unified Medical Language System (UMLS) thereby creating a reproducible and transferable crosswalk to existing Metathesaurus concepts. The corpus of CC data was taken from the EMRs of three southeastern US EDs representing urban, rural and suburban academic medical centers. There were 39,038 patient visits and 13,494 unique CC entries examined in the corpus. An unclear method of exact concept matching followed by normalized matching was used.

Further processing using a combination of automated and manual techniques by a single expert ED nurse reviewer was utilized. The expert reviewer performed frequency counts and tokenization. Matched UMLS concepts after round one pre-processing was 14%. NLP routines based on the developed algorithms were written as Pearl programming language scripts. Three groups of NLP routines were developed and applied in successive rounds using a simple to aggressive technique escalation. First, they identified commonly used punctuation patterns such as slashes, commas, and semi-colons and removed them during pre-processing of the data.

Matched UMLS concepts after pre-processing, and round one processing was 9%. A second level of processing analyzed unmatched entries from the first round and

processed the data to address acronyms, abbreviations and truncated words. Matched UMLS concepts after round two processing was 7%. In the final round, unmatched entries that remained from the second round of processing were processed to address modifiers and qualifiers. Matched UMLS concepts after round three processing was 18%. Overall, 86% of the matched entries were identified with one UMLS concept only, and 14% more were identified with two or more UMLS concepts. Attempting to match CC, which is written in free text to a standardized ontology, looks to be a difficult one. In general, this method is time-consuming and inaccurate. Multiple levels of processing are necessary to remove the unique documentation and punctuation which are specific to ED clinical notes. The results of this work are less important than the thought process used to complete the work itself. This article demonstrates the real need for clinicians to have definitions that are reliably linked to free text sections of the clinical record in the EMR. This gap has been recognized by the nursing profession in this work and supports the need for a tool. In this particular example, the tool utilized was not the correct one. However, the question which was attempted to be answered was a telling one.

   The work by McPeek-Hinz, Bastarache, and Denny at Vanderbilt University Medical Center adds significantly to the best practice for identifying venous thromboembolism (VTE), both acute and historical, in electronic health records (EHRs) using both ICD-9 codes and natural language processing (NLP).[106] The authors investigated an NLP processing algorithm to define a VTE phenotype. Using ICD-9 codes alone, VTE could be poorly predicted (PPV = 0.29) with fair sensitivity (0.68). AHRQ Patient Safety Indicator algorithms when used for VTE identification yield slightly better results (PPV = 0.545, Sensitivity 0.87). The AHRQ definition can be found in the definitions section of this work. In this study, the authors used both ICD-9 and NLP methods to identify VTE

cases in a retrospective review of 9,504 patients registered in the Vanderbilt DNA

biobank (BioVU). This work provides detailed definitions used to identify VTE which

usually is under-reported in the literature and can be found in the definitions section of

this work. The study used two board-certified physicians to review the findings in order

to confirm all free-text identified possible "hits" for positive (N=590) and negative

patients (N=8914) as defined by the algorithm. The significant amount of time spent by

these reviewers was not reported in this work. However, using ten minutes per chart for

extraction, as reported by Gehrman, Dernoncourt, Li, et al., an estimated amount of time

could reasonably be assumed to be 740 hours per reviewer to review each negative case

without an ICD-9 positive "hit."

The NLP, which the authors used, was the KnowledgeMap Concept Identifier (KMCI),

which is a general-purpose NLP program proprietary to the Vanderbilt medical center.

The NLP algorithm had difficulty and required multiple attempts to control for false

positives. Examples include hits for "PE" for "physical exam" instead of Pulmonary

Embolism and a patient being "at-risk" for a DVT, a "possible complication" of surgery,

or a patient needing DVT prophylaxis. The authors utilized the NLP algorithm to

examine for the presence of VTE using three data sets compared against the physician

reviewed gold standards (PPV = 0.69). The learned NLP algorithm was used to identify

VTE presence in patient problem lists (PPV = 0.972, Sensitivity = 0.42.8, F-measure

0.594), clinical notes (PPV = 0.91, Sensitivity = 0.916, F-measure 0.908) and both

problem lists with clinical notes (PPV = 0.90, Sensitivity = 0.951, F-measure 0.925). The

conclusion is that VTE disease identification is performed best when derived using ICD-9

codes and further refining using NLP processing of both hospital notes and problem lists

of physicians. This article is important due to the specific definitions given for training

NLP processes and the presence of VTE in a logical clinical context. The findings are not

able to be replicated outside of this institution secondary to the proprietary database tools developed and deployed in this specific research.

Machine Learning

Kotfila and Uzuner examined the effect of doubling the training set data on feature spaces, feature weights, and support vector machine (SVM) kernels on model phenotyping performance.[107] The authors used both the 2008 and 2014 i2b2/UTHealth database to investigate five diagnostic groups: obesity, atherosclerotic cardiovascular disease, hypertension, hyperlipidemia, and diabetes. Pre-processing of data prior to NLP use reported by the authors in this work were: Minimally normalized tokens were generated by preprocessing the raw text of documents to remove numbers and non-alphabetical characters, text was converted to lower case, common stop words (127) were removed, tokens were split based on sequences of one or more contiguous white space characters, the K best features were selected using a univariate parametric filter based on a one way ANOVA F-test, and classification was performed with two different SVM kernels: a linear kernel and a Gaussian radial basis function (RBF) kernel. The authors used the LuiNorm tool, MetaMap, Standford Core NLP sentence breaker, and NegEx to perform this work. The authors found that the addition of training data has a weak effect and weak statistical significance across disease classes. This is important to note because the increased size of training corpora does not necessarily lead to increased model performance. The authors confirm this in stating that, "For each disease, the threshold for the level of training data varied with each disease. The initial conditions under which the corpus is designed can affect model performance and should be considered when comparing techniques across different corpora."

Beaulieu-Jones and Greene deployed semi-supervised learning of the electronic health record using denoising autoencoders for Amyotrophic Lateral Sclerosis (ALS) phenotype stratification.[108] The author's primary tool in this work was the use of Denoising autoencoders (DAs) to perform unsupervised machine learning. DAs are a type of artificial neural network trained to reconstruct an original input from an intentionally corrupted input. Through this training, they learn higher-level representations modeling the structure of the underlying data. The DAs were combined with a classification algorithm, random forests, and found improved survival prediction in ALS clinical trial data. The results of DA predictive analysis closely mirrored the gold standard of support vector machine predictions. DAs are an unsupervised alternative for phenotype modeling using EMR data. DAs are relatively inexpensive to perform analysis on large amounts of unlabeled EMR data. Since it is expensive to have data labeled for use in research by a clinician, DAs make an excellent choice for reducing this associated cost. DAs in association with a random forest algorithm demonstrate performance on par with SVMs when there are many unlabeled samples and few labeled samples. This work supports the fact that EMR-phenotyping is extremely useful in identifying cases where only a small number of patients belong to a phenotype. It also lends support for possible future utilization of this method to expand the anchor and learn the concept from semi-supervised to unsupervised performance.

<div align="center">SUPPORT VECTOR MACHINES</div>

Lin, Hsu, Lou, et al. demonstrate that support vector machines are proven to have the best performance in free-text medical writing classification, compared with Naïve Bayes classifiers, C4.5 decision trees, and adaptive boosting.[109] This work details some little reported information in this type of research. Namely, the equipment specifications

used for their machine learning and artificial intelligence processing. The authors used a

Fujitsu RX2540MI, 48 core CPU with 768 GB RAM server. The all-flash array used was

an AccelStor NeoSapphire, NS3505 with a 5 TB serial advanced technology attachment-

interface solid-state drive and connectivity of 56 GB/second FDR InfiniBand Quad Small

Form-factor Pluggable. This is a wholly overlooked piece of information that has slipped

by the peer review process in published journals. In such journals as AMIA or the

Journal of Biomedical Informatics, I assert that equipment utilized should be a standard

reported item as many methods are being examined and deployed in novel ways during

this growth phase within the biomedical informatics discipline. This work also

demonstrated the author's successful use of Word visualizations in order to demonstrate

the differences between Training and Testing set values used. This tool seems to have

good applicability for generalized physician communication of this subject matter.

NATURAL LANGUAGE PROCESSING

Mehrabi, Schmidt, Waters, et al. used natural language processing (NLP) to identify

pancreatic cyst identification in electronic health records using the Refenstrief institute

database of clinical data.[110] What was novel about this work was that accuracy in

identifying targeted clinical records was increased by prefiltering syntax before negation

detection. The authors used the Stanford data parser (SDP) before running the NegEx

algorithm. The SDP analyzes how words are related to each other in a sentence. NegEx

is an algorithm that looks for negation terms like "Rule out, No evidence of" within a

targeted sentence. The identification of false negatives in written language is essential to

the increased accuracy when processing via NLP algorithms. The need to process the

meaning of sentences adds a significant burden to machine learning through the use of

NLP when using clinical free text. Clinical documentation does not follow traditional

sentence structure in the English language and therefore creates unique difficulties using traditional NLP methods.

The work by Zhou, Baughman, Lei, et al. demonstrates that such obscure and elusive diagnoses such as depression can benefit from the use of the combination of natural language processing (NLP) and machine learning (ML).[111] The authors were able to successfully identify hospitalized, ischemic heart disease patients with a depression diagnosis utilizing NLP pre-processing and ML techniques by a factor of 20%. Complex, multimodal diagnoses such as depression were able to be detected with the inclusion of free text. The accuracy improvement of 20% was not the critical finding in my opinion. The ability to use NLP pre-processing that increased accuracy lends support to the use of free-text analysis of patient notes in predictive work.

The addition of free-text clinical notes continues to show up in the literature as a best practice for increasing clinical predictive accuracy. Kontio, Airola, Phikkala, et al. looked at a novel method of predicting patient acuity from nursing notes within electronic medical records (EMR).[112] They used the regularized least-squares (RLS) or the Ridge regression machine learning (ML) model to predict what degree the clinical information in the EHRs of cardiac patients can be used to predict patient acuity scores for the following day. The methods used were based on linguistic pre-processing, vector-space modeling of the text and regularized least squares regression. The results show that the ML learning approach is significantly better than simple approaches such as predicting today's score from the previous day or a majority score. This score was incrementally improved by adding in nursing free text information and historical previous acuity score data. This was further improved by the availability of real-time data access to notes from the same day. In this work, the application is one of carryover. The ability of a machine

learning model to predict a future state, from a former state within a clinical domain is exciting. The ability of nurses' free text within the EMR to increase the accuracy of the model demonstrates yet more support for the use of free text as a needed expansion of knowledge beyond categorical and Boolean variable use for prediction.

Central Repositories

Jiang, Kiefer, Rasmussen, et al. demonstrate that data does not exist, or significant gaps are present when trying to gather data element repositories for phenotyping and interoperability.[113] Many research communities have formed to attempt to solve this problem. The most commonly cited and utilized attempts have been put in place by the Electronic Medical Records and genomics (eMERGE) network, the strategic Health information technology advanced research project (SHARP), The HMO research network (HMORN) and the National patient-centered clinical research network (pCORnet). See the definitions section of this work for further information on each of these networks. In their work, Jiang et al. propose yet another model called the phenotype execution and modeling architecture (PhEMA). The objective of the research is to develop and evaluate a data element repository (DER). The DER is to provide standardized representations and machine-readable application programming interfaces (APIs). The DER is to use the ISO/IEC 1179 International metadata standardization. The existence of the ISO standard is a significant step forward for such repositories. The authors have chosen the Quality Data Model (QDM) as an information model for representing phenotype algorithms. The QDM was developed in 1999 by the National Quality Forum (NQF) for representing EHR-based electronic clinical quality measures (eCQMs). The QDM specification is available only as descriptive text documents. Therefore, they require human interpretation for broader use and analysis via machine

learning methods. There are several underspecified areas within QDM despite it's scalability and the simplicity of the DER methods proposed. The challenge for the future for all clinical repositories is one of interoperability, complete end to end machine-readable code, and most importantly, adoption by both public and private members of the repository.

Banda, Halpern, Sontag, et al. investigate electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network.[114] The authors in this work present Automated Phenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE). Using the R programming language, they combine noisy labeling and anchor learning to allow for phenotype use with the OHDSI data network. The OHDSI collaborative has 140 collaborators in 16 countries with a shared vision of improving community health through evidence-based medical practice. The OHDSI has a standard vocabulary and a common data model (CDM). These tools allow for systematic analysis of disparate observational databases from its members. The CDM stores EHR data, claims data and a standardized vocabulary of over 650 million patients. Since it is coded to use the open-source R-programming language, it is open for researchers to take advantage of phenotyping via supervised or semi-supervised learning of phenotype models. APHRODITE is based upon learning with noisy labels theory presented by Agarwal, Podchiyska, Banda, et al. and expanded by Halpern, Choi, Horng, and Sontag in their Anchor and Learn framework. In this published work, the authors compare their performance results when building phenotype models for Myocardial Infarction and Type 2 Diabetes Mellitus using four different models. The APHRODITE method yielded a superior mean demonstration accuracy (0.93) and positive predictive value (0.91) for both phenotype models. In this

work, the authors successfully show that it is possible to identify anchors during the learning of a dataset model using a standardized common data model from the OHDSI data network.

Mate, Castellanos, Ganslandt, et al. perform a proof of concept for standards-based procedural phenotyping using the Arden Syntax on the Informatics for Integrating Biology and the Bedside (i2b2) dataset.[115] The focus of this work was to demonstrate that a complex phenotype could be constructed and deployed using a standard programming tool and visual methods. The Arden Syntax for Medical Logic Systems is an HL7 standard designed to enable clinical decision support functions in the form of Medical Logic Modules. It was intended to be used as a tool for those without computer science degrees. The authors designed a Java-based code option to create a visual tool that accessed the data from the i2b2 dataset for phenotype creation. The authors did succeed in demonstrating the functionality of their tool using the Arden Syntax Medical Logic Module concept. There were many potential drawbacks to moving forward with this approach. Currently, only large institutions have the necessary computer science support to deploy and utilize the Arden Syntax programming language. Second, a significant workaround was needed due to a version change in the Arden Syntax software, which hampered the function of the newly created visual tool. This work demonstrates the difficulty in using multiple sources of the software in a manner for which they were not created.

VENOUS THROMBOEMBOLISM

Lecumberri, Panizo, Gomez-Guiu, et al. examined the economic impact of an electronic alert system within the EMR to prevent thromboembolism in hospitalized patients.[116] The authors linked an electronic alert (e-alert) software to the patient database in acute care, and Joint Commission accredited hospital in Pamplona, Spain. Using medical orders, daily nursing reports, surgery registries, and lab results, a patient's thrombotic risk was calculated. Over the five years from 2005-2009, half of the VTE cases (51.8%) were surgical patients, with the rest being medical patients. The median length of stay was four days across the five years. The mean age in years was 55.4, with 20% being higher than age 70. A new phenomenon presented itself during the review of the data. The physicians were more likely to follow appropriate prophylaxis recommendations in surgical patients (85% adherence) Vs. Medical patients (60% adherence). The percentage of all medical patients in which an alert was sent averaged 14.3% over the five years. Whereas, 40.9% of the surgical patients had an alert sent. This was a new bias that was discovered by the researchers. Physicians were more likely to have an e-alert sent by the automated system and to respond appropriately to that alert when patients were post-surgical Vs. Being treated for general medical conditions. The authors went on to then tie in the costs related to this research. The mean direct cost (during hospitalization and after discharge) of an in-hospital VTE episode was found to be $10,234. Direct costs per single hospitalized patient were reduced after e-alerts from $31.30 to $17.10, while the increased use of thromboprophylaxis and the

development of e-alerts meant $4.35 and $0.51 per patient, respectively. Thus, the implementation of e-alerts led to a net cost saving of $9.40 per hospitalized patient.

This research demonstrates a few important considerations. First, automated CDSS tools cannot be considered a reliable method for shaping clinical behavior. Second, if the identified physician bias towards post-surgical identification of VTE versus those with general medical issues carries over into a gross generalization for physician clinical practice, then a genuine need exists for tools that identify the potential risks for other healthcare professionals. This might be most readily seen in an example within the population most at risk for VTE who reside within the assisted and extended long term care facilities within the United States. Given that the patient population within nursing homes contains both postoperative and medical patients, the physical therapist can not rely on primary physician screening on medical patients. If physical therapist behavior is the same as physician behavior, then PT's may be at risk for decreased identification of VTE in medical patients as well. In this case, CDSS tools would prove useful.

Umscheid, Hanish, Chittams, et al. at the University of Pennsylvania used a CDS intervention to assist in VTE prophylaxis.[117] It was unique due to the lack of pop up alerts in order to eliminate provider alert fatigue. The study included 223,062 inpatients across three hospitals, including 1,714 beds in the study period between April 2007 and May 2010. All facilities used the Allscripts EHR platform. An interdisciplinary team composed of physicians, nurses, quality specialists, pharmacists, informatics analysts, and anticoagulation experts designed the CDS tool. The CDS tool required the admitting provider to accept or decline VTE prophylaxis based on suggested patient risk. Low risk is the absence of one of 11 risk factors. (1. Age ≥40, 2. recent surgery lasting ≥ 45 minutes, 3. history of venous thromboembolism, 4. history of hypercoagulability, 5.

history of cancer, 6. obesity (BMI ≥30), 7. ongoing estrogen or antiandrogen use, 8. History of varicose veins, 9. reduced mobility, 10. weakness or paralysis of ≥ one limb, 11. expected length of stay three days) The CDS did not auto-populate risk factors based upon presence within the chart. Providers were then asked to select one of three options: 1) pharmacologic prophylaxis; 2) mechanical prophylaxis only; or 3) no prophylaxis. Reasons were required if either of the first two options was selected.

VTE events were defined as any hospital discharge with a secondary discharge diagnosis of PE or DVT as defined by the International Classification of Diseases Version 9 (ICD9) codes listed in the Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSI) Technical Specifications Guide under PSI 12.[118]

The results of the study show that "recommended" prophylaxis significantly increased across all hospitals and services (27.1% vs. 51.9%; p<0.01). "Other surgical services" had the most considerable increase in recommended prophylaxis overall (19.8% to 48.2%; p<0.01).The orthopedics/trauma services had the lowest rate of increase overall (44.4% to 48.8%; p<0.01), but had the highest rate of "recommended prophylaxis" before the CDS interventions. "Pharmacologic" prophylaxis also increased across each of the three hospitals and the health system overall (42.0% vs. 54.4%;p<0.01). "Other surgical services" had the highest increase in pharmacologic prophylaxis overall (33.3% to 49.2%; p<0.01). In general, there was a marked reduction in "no prophylaxis" for all services, most prominently orthopedics/trauma, which decreased its overall "no prophylaxis" rate from 34.1% to 3.0% across the three study periods (p<0.01).

Increased awareness of an emphasis on VTE prophylaxis in this subpopulation as a result of the AHRQ PSI metric may have caused the more significant improvements in VTE prophylaxis demonstrated in this subpopulation. Specifically, those services

described as the "other surgical services" in the study had the highest increases overall in both recommended prophylaxis (with an approximate absolute increase of 30% over the study period) and pharmacologic prophylaxis (with an approximate absolute increase of 20% over the study period) intervention may have more impact in populations at higher risk of VTE (such as the surgical population defined by the AHRQ measure). Many hospitals use Allscripts; however, this type of administrative review would not be able to be universally applied to other institutions using alternate EHRs.

These findings are in line with that of Lecumberri, Panizo, Gomex-Guiu, et al., who also demonstrated a lack of support for VTE identification and prophylaxis by physicians when patients were post-surgical Vs. being treated for general medical conditions[116]. However, this study demonstrates that behavior can be significantly influenced to come into compliance with evidence-based practice guidelines when a CDS is deployed in a minimally invasive way.

Kucher, Puck, Blaser, et al. demonstrated that EAlerts show promise in reducing the rate of VTE secondary to increased clinician awareness of prophylaxis.[119] The department of medicine at the University Hospital of Zurich, Switzerland, turned on such an eAlert system for all acute care patients, not just those at risk for VTE for 14 months, including the period from September 2007 to December 2008. The eAlert was set to continuously flash until primary physician attention was given to the alert. Physicians needed to select one of two options, prophylaxis indicated, or prophylaxis not indicated. Appropriate prophylaxis included either pharmacologic or mechanical methods. The eAlerts went off 6 hours after admission if the prophylaxis was not ordered regardless of the indication VTE prophylaxis was present or not. The eAlert was visible to all healthcare professionals involved with the case. Automatic switching off of the

alert was at discharge, a transfer, or at ten days whichever came first. A classic case of

alert fatigue was demonstrated by the physicians who cared for > 20 patients during the

study period. These physicians not only answered the alert more quickly but had a lower

rate of appropriate prophylaxis within 6 hours after admission (37% Vs. 50%). Despite

this anomaly, the rate of appropriate prophylaxis among hospitalized patients increased

from 44% to 76% in this institution.

Clinical Practice Guidelines

PRACTICE

Goldbraich, Waks, Farkash, et al. looked at computational natural language

processing (NLP) methods for reviewing deviations of oncology physicians from clinical

practice guidelines (CPGs), as noted in patient discharge notes.[120] Approximately half

(48.9%) of all the treatment patterns deviated from the CPGs. The most significant

reason for deviation was a tendency to overtreat patients. This was 3.1 times greater in

frequency than the next closest reason for CPG deviation which was missing treatments.

NLP has been used to examine CPG deviation in discharge summaries within an

electronic health record's free text with excellent results.

Planquette, Maurice, Peron, et al. The lack of knowledge of common clinical practice

guidelines is not confined to professions outside of medicine.[121] Planquette et al.

surveyed general practitioner (GP) physicians in private practice in Paris France via

questionnaire and two clinical cases with an excellent return rate. (30.5%). In general,

most GPs were not aware of the diagnostic algorithm for pulmonary embolism (PE).

Specific training on PE via continuing education and knowledge of clinical probability

scores were positively associated with the use of the validated PE algorithm. A majority

were not able to link the clinical probability of PE and D-dimer testing. In fact, over 80% did not correctly identify the clinical purpose of D-dimer testing in this patient population. The authors hold that GPs are most often the first line of assessment contact for patients with non-critical cases of pulmonary embolism. GPs are less likely to use guidelines than hospital-based physicians.[122] This work continues to show the need for automated, risk-based clinical decision support tool used for all healthcare clinicians.

North, Fox and Chaudhry Computer-assisted pre-processing of data is the promise of the future of medicine.[123] Having information coded in a computer-readable format allows for automatic presentation of time-saving tools was the topic of research by North et al. North has pointed out that cardiac physician provider times to calculate related cardiac risk scores, even with electronic tools, is too burdensome for a full caseload of patients. The author suggests that in order to increase efficiency in workflow and therefore make the best use of available time in the clinic by physicians, natural language processing and ML algorithms are needed to "serve up" automatically calculated risk assessments to physicians at the point of care. This supports the direction of my work to include triggers that fire automatic CDSS tools for clinicians. In this case, physicians would be the benefactors; however, other professions such as physical therapy would benefit from such technologies.

Summary

This review of the literature contains several themes that support the need for computer-assisted phenotyping. The lack of knowledge of common clinical practice guidelines is not confined to professions outside of medicine. Clinical decision support tools would prove useful and have been shown to significantly influence compliance with

evidence-based guidelines when deployed in a minimally invasive way. Phenotyping of patients is a necessary next step to allow for these algorithms to automatically serve up calculated recommendations to clinicians at the point of care. The development of comprehensive phenotypes is a needed next step towards comprehensive, evidence-based practice.

The manual extraction of rule-based, algorithmic identification of phenotype phrases is too time-consuming to be effective. Proprietary database tools and curated centralized data warehouses are not the solutions with their high barriers to entry and operation. Well documented, open-source tools are required as an alternative. Without an identified standard method of phenotype development, industry network workgroups must shift their focus towards tool analysis instead of curation of data libraries. The performance of these tools should be reported on and discussed within the literature before a new tool is developed.

Systems that make holistic use of the electronic medical record in characterizing a patient for phenotyping is needed. For these needs to be met, machine learning tools in the form of support vector machines and natural language processing must be utilized. Efforts to allow phenotyping algorithms to focus and scale are needed to be reported in the literature. Having the ability to predict the presence of a disease state early may lead to optimal patient outcomes, decreased morbidity, and prolonged quality of life.

Chapter III

METHODS

Overview

Methods for performing anchor phenotype identification as previously described in the literature by Halpern, Choi, Horng, and Sontag.[21, 90] were followed to reproduce, as well as, validate the published protocol. These methods were modified to shift the focus from physician-centric documentation free text sources, to a dataset comprised of physical therapy free text sources. A single-subject feasibility analysis was undertaken before full data access and solidifying final methods.

Single Subject Feasibility Analysis

The data holding facility, Accord Care Community (ACC), has a history of excellence and has been awarded a five-star rating by Medicare. This achievement is attributed to the ability of the facility to gather, analyze and make informed clinical patient outcome decisions based on data. Secondary data was readily available for review and utilized in this research. The ownership of the facility and this researcher had previously worked together in the post-acute care environment thereby establishing the required

professional and ethical screening required for this level of collaboration. The facility

was initially concerned about patient privacy and confidentiality. Multiple discussions

were had with the facility, facility legal council, chair of the IRB at the University of

Mount Union and the chair of this researcher's dissertation committee. A final agreed-

upon course of action required a HIPAA Business Associate Agreement to be signed by

the principal investigator with the ACC and was executed in January 2019. A copy can

be viewed in Appendix D.

This researcher also developed and presented a data safety and migration plan to the

ACC in January 2019. This plan was accepted by the owner of the facility, the director of

rehabilitation, legal counsel of the facility and by the Ohio Healthcare Association.

Approval from an Institutional Review Board (IRB) was not a requeirement, however it

was pursued as part of faculty governance requeirements at the University of Mount

Union. A separate data use agreement governs access to the data in this research. A copy

of the Data Use Agreement may be viewed in Appendix E. This plan was communicated

to the chair of this dissertation committee with approval. The study was approved by the

University of Mount Union's Institutional Review Board, Committee on Clinical

Investigations Protocol #UMU_IRB_353. A waiver of informed consent and

authorization was granted in April 2019 by the Committee on Clinical Investigation, as

described in 45 CFR 46.116(d). A copy can be found in Appendix F.

Since a single patient design was used for this validation stage of the research, the

published methods were modified. First, the "Anchor elicitation tool" was not utilized.

Second, the facility contact at the ACC acted as the physical therapy domain expert

reviewer to validate key conditions. This researcher acted as the domain expert to

validate all text anchors. Free text fields were extracted by hand from the physical

therapy documents, as described in the detailed method data. An initial review

preparatory to research was performed on a single patient record with an admitting and treatment diagnosis of deep venous thrombosis. This researcher was given a single facility contact, the director of rehabilitation, who extracted the necessary information required to allow for analysis of a single patient record. The patient was selected at random by the facility contact from all positive cases of DVT via ICD-9/10 diagnosis coding in PointClick Care. The patient selected had both an admitting and treatment diagnosis of deep vein thrombosis. Data were collected and processed as per the methods described in the detailed method data preparation section in this work. Third, Machine learning was not necessary to train and predict a learner using this method.

Finally, the anchor weights calculation was substituted for feature weights in the free text analysis. Feature weights were estimated using text pre-processing and resultant bag-of-word creation and frequency weights using the following method. A separate corpus was created for each parallel physician matched grouping. Table 8 details these

TABLE 8. Matched Free Text Groupings Between Halpern et. al. Physician data and Proposed Volansky Physical Therapy Data.	
Physician*	**Physical Therapist**
Physician Comments	Certification Evaluation • Past Medical History • Objective Tests/Other Precautions • Other Comments/Observations
	Discharge Summary • Other Comments
Triage Assessment	Therapy Progress Report • Patient Response to Treatment This Week
* Halpern Y, Choi Y, Horng S, Sontag D. Using Anchors to Estimate Clinical State without Labeled Data. AMIA Annu Symp Proc. 2014 November 14,;2014:606-15.	

matched free text groupings. A corpus was created from the physical therapy initial certification and discharge documentation. A second corpus was created from the progress note documentation. Each corpus was pre-processed to delete common words, punctuation, capitalization and blank space. Each corpus was processed into unigram,

bigram, and trigram tokens appearing in the bag-of-words. Frequency counts were

calculated utilizing a simple open-source tool available at

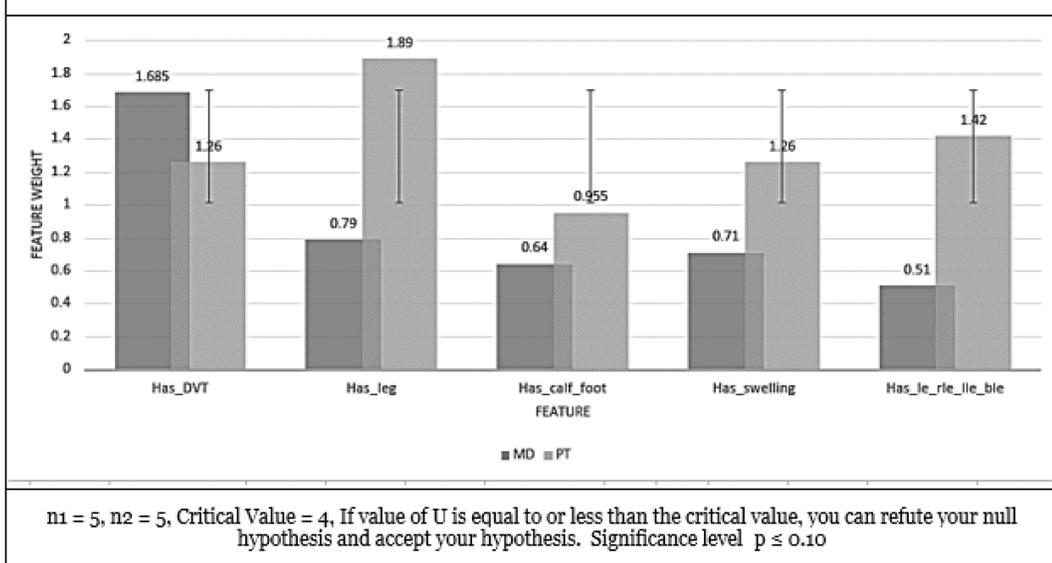http://guidetodatamining.com/ngramAnalyzer/index.php. Both of the corpora were

TABLE 9. nGram Description of Free Text Corpus.

nGram	Corpus 1 (Certification + Discharge)		Corpus 2 (Progress Report)		Corpus 3 (Corpus 1 + Corpus 2)	
	Tokens	Types	Token	Types	Tokens	Types
Unigram	318	198	462	116	779	277
Bigram	318	272	462	219	779	489
Trigram	318	288	462	271	779	559

then combined into a single bag-of-words and reprocessed with an updated frequency

count for the final combined corpus and can be viewed in Table 9. Frequency counts

were reported and compared to the feature weights reported in the literature by Halpern

et al.

Results of this initial feasibility work, as shown in Figure 7, reveal a subset of a

statistically different set of feature weights when physical therapy clinical documentation

was used for feature extraction versus emergency physician notes. (Mann Whitney U-

test, p ≤ 0.10) This feasibility work demonstrates that a gap does exist in the free text

sections of DVT patients when physical therapy data is used in comparison to physician

data. This information was shared in the Rutgers Biomedical Informatics Spring

Colloquium presentation of this work on April 25, 2019. There were no comments nor

questions which arose during the presentation or offline in follow up to the colloquium

presentation from faculty and students who attended.

FIGURE 7. Common Feature Weights Physician v. Physical Therapist.

n1 = 5, n2 = 5, Critical Value = 4, If value of U is equal to or less than the critical value, you can refute your null hypothesis and accept your hypothesis. Significance level p ≤ 0.10

Methods

Data Origin

The examined data set was compiled from an 82 bed, for-profit, Medicare five star rated and certified skilled nursing center in Middleburg Heights, Ohio, Accord Care Community (ACC). (https://www.accordcc.com/) A retrospective analysis of all patients receiving skilled physical therapy care during the six years between 2012 and 2018 was performed using secondary patient data. Initial data was compiled for quality improvement and performance improvement efforts and made available for use in this research.

Data Parity

Several modifications were necessary, which deviate from the described methods. First, the elimination of the collection and use of the variables described as dispensed medications, lab values and triage vital signs. These variables were included in the original work to help validate changes in clinical condition over time spent in the emergency department. These variables are not of use given that time series data is not the focus of this research.

Second, the utilization of the medication history data was not eligible for inclusion in the study. The ACC electronic medication administration record (eMAR) was initially been reported to be present for all episodes on or before the survey inclusion dates. After closer review, only 18 months of data were available in a format that could be utilized. The remainder of the eMAR data was either archived in a proprietary format or only available via paper record.

Data Preparation

The data required extensive pre-processing after being extracted from two proprietary electronic medical record databases used by ACC, Point Click Care, and Therapute. Each database required a different strategy and, therefore a different methodology for its extraction and pre-processing. All relational database queries were made and exported into. CSV readable format in Excel. Where data was unavailable for export directly to a relational format, individual reports were constructed, and data were extracted and formatted manually from .pdf or .doc formatted reports using Excel.

Data was transferred and stored via the Box cloud storage service. (Available at https://www.box.com) The service plan was upgraded to "Box for Healthcare," which allows for HIPAA compliance, 256-bit encryption of data in transit, and at rest and for audit purposes.

Data safety included the following measures to mitigate the risk of a breach. All data associated with this research was stored in a secure cloud storage location. All data transferred to and from this location used VPN encryption provided by https and via the Tunnel Bear private VPN service (https://www.tunnelbear.com). Database cloud access was restricted to two (2) people (This researcher and the Facility Contact at ACC). All passwords utilized in any application or hardware associated with this research were complex (upper case, lower case, numbers, special characters) using 20 characters and 128-bit encryption. Data access on this researcher's computer required a biometric scan (finger) in addition to a complex password. Data was stored on an external hard drive when locally manipulated and was 128 bit encrypted with secure password use. Upon completion of the research, the external hard drive containing the local database was deleted via the Department of Defense protocol (DoD 5220.22-M) using the open-source Eraser software (Available at https://sourceforge.net/projects/eraser/).

All patient-level data collected were de-identified following removal of the 18 recommended fields as listed in the HIPAA patient data de-identification "Safe harbor" protocol.[43] The HIPAA Rule for Anonymization can be found in the definition section of this work. Patient Identification numbers assigned randomly by the Point Click Care (PCC) EMR were used as primary keys to connect episodic data with the Therapute (TP) EMR. Episode start dates were appended to each patient identification number to mark single episodes of care. Each episode represents a single patient record. No patients

were excluded, thereby leading to a total of 1,668 complete patient encounters for analysis.

Patient records are represented as containing three distinct types of observable variables that come from structured sections of the PCC EMR: 1. ICD9 and/or ICD10 diagnosis codes, 2. Age, 3. Sex. Latent variables which come from the free text sections of the TP EMR are formed by a concatenation of the PT Certification Document (from the 'past medical history', 'objective tests/other precautions', and 'other comments/observations' sections), the PT Discharge Document (from the 'other comments' section) and the PT Progress Report (from the 'patient response to treatment this week' section).

Point Click Care Electronic Medical Record

ACC utilizes point Click Care (PCC) for billing and for documentation of clinical nursing services, which include medication history. Multiple summary level reports were run for all episodes starting on or after

TABLE 10. Data Features Used to Build Binary Feature Vectors.			
Data Feature	**Representation**	**Dimension**	**Source**
Age	Binned by decade (20-90)	8	Point Click Care
Sex	Male / Female	2	
ICD-9-CM / ICD-10-CM	ICD-10 used on or after October 1, 2015	2156	
Certification Evaluation	Binary Bag of Words		
Discharge Summary	Binary Bag of Words	9653	Therapute
Therapy Progress Report			
Total Feature Dimension = 11,819			

January 1, 2012, and ending on or before December 31, 2018. The final data features used to build binary patient description vectors are Age, Sex, ICD-9-CM and/or ICD-10-

CM code are reviewed in Table 10. Conversion of ICD-9-CM to ICD-10-CM codes was performed utilizing a combination of manual and batch processing via the conversion tools available from the American Academy of Professional Coders. (Available at https://www.aapc.com) Where specific conversions were not directly possible, manual selection of appropriate generic equivalence mappings was used.

Therapute Electronic Medical Record

Therapute (TP) is utilized for administrative, billing, and documentation of clinical physical, occupational and speech therapy services. Clinical free text data was gathered from several documents and locations within TP. The parallel data to the Emergency Physician notes free text were found in two places. The first location for the free text data is the Physical Therapy Certification Note (from the Past Medical History, Objective tests/other precautions, Other comments/observations sections.) The second location for the free text data is the Physical Therapy Discharge Note (from the Other comments section.) The Emergency Triage notes free text was found in the Physical Therapy Progress Report in the area of patient response to treatment this week. All reports were exported into.CSV format before preprocessing.

Representation and preprocessing

Deidentified free text was preprocessed using bigram and negation detection before being represented as a binary bag-of-words (BoW). Examples of preprocessing include the removal of punctuation, capitalization, formatting, proper name and identifiable data (age, gender, birth dates). This level of processing was done using the Porter Stemmer

function (Available at https://tartarus.org/martin/PorterStemmer/) and Matlab.

Negation phrasing took place using "negex" rules [124] with some manual adaptations appropriate for physical therapy notes and negated words were replaced by a new token (n.b. if the token "nonhealing" was within the scope of negation, it was transformed to a new token, "neg_nonhealing"). A second step of preprocessing collected 250 significant physical therapy bigrams and appended them to the text. For example, if the phrase "calf pain" was augmented to be "calf pain calf_pain" with an extra token representing the bigram). When learning with anchors, the component words are removed (i.e., "calf pain" is replaced by a single token "calf_pain." This was done in order to increase the amount of conditional independence between anchors, which are bigrams and the rest of the text. If the token "calf_pain" is chosen as an anchor, it will not be conditionally independent of the tokens "calf" and "pain" without the removal step.

The binary BoW was processed using Token Frequency - Inverse Document Frequency (TF-IDF) processing in Matlab with resultant creation of a binary feature vector. Logistic regression was then utilized to identify related classifiers for venous thromboembolism anchor specification

Anchor Specification

Anchor terms identified from the logistic regression were cross-validated by a single physical therapist content expert who specified anchors for each of 139 final clinical state variables using a five-point Likert scale. Scores of 4 and 5 were taken to be positive and everything else to be negative. The survey was administered, and results collected electronically using the Survey Monkey platform. A copy of the survey can be found in Appendix G. A single content expert was assigned based on their following qualifications

for inclusion in the study. They were identified as having practiced cardiopulmonary physical therapy (≥ 25 years) and taught cardiopulmonary physical therapy (≥ 20 years). Note that the physical therapist was not be provided with explicit feedback about the performance of the model. In order to accurately assess the method involved in specifying anchors for a new classification task via the Survey Monkey platform, a second content expert was asked to repeat the survey for validity reasons only. The second content expert was assigned based on their following qualifications for inclusion in the study. They were identified as having practiced physical therapy (≥ 25 years), taught physical therapy (≥ 20 years), and was an orthopedic clinical specialist (≥20 years). The interrater agreement between the two experts using the survey tool achieved a kappa value of 0.8430. A final concatenated feature vector with 9,663 fields was generated for analysis.

Mulitnomial Logistic Regression Modeling

The learned feature vector was used to build a multinomial logistic regression model (MNR) using Matlab. The learned MNR model was used on the complete held database of all patients who received skilled physical therapy care during the six years 2012-2018 to predict the presence of the VTE phenotype. A comparison was made between the classifiers learned using anchors to a simple rule-based baseline and a supervised MNR baseline. An evaluation was reported on separate estimation tasks, one for each clinical state variable identified in the Anchor specification phase.

The rule-based baseline used ICD-10-CM codes as a reference for positive VTE cases associated with an anchor term. The rule-based baseline predicts positively when at

least one anchor is present and negatively otherwise. This approach was evaluated on the entire labeled dataset.

Based on the sparsity of the dataset, evaluation of the supervised baseline was completed using randomized samples. In each experiment, random samples of both the labeled (N=30) and unlabeled (N=30) patients were selected. This selection process was repeated 500 times for each of the three individual anchor test sets with a resultant feature size of 30,000 random samples. The results of the selection were analyzed using Matlab for an estimate of the nominal responses and weighted for statistical significance. The absolute values of the coefficient estimates were used for comparison and frequency counting. Single feature tokens, which were found to have zero examples across the prediction runs, were dropped from the analysis. The overall performance of the model was measured using the rank ordering of anchor presence in the randomly selected test set.

Data Classification

Data classification analysis was performed using three individual corpora of free text derived VTE phenotype estimators. These estimators were compiled from Expert Selected Anchors, Anchors extracted from the Academy of Acute Care Physical Therapy Clinical Practice Guidelines (AACPT CPG), and anchors extracted from both labeled and unlabeled dataset. The expert selected anchors were selected based on the Likert scale score. The anchors selected from the AACPT CPG were derived using natural language processing of the entire CPG published corpus using the same methods described in the representation and preprocessing of free-text data. A binary BoW was created with significant tokens processed into a TF-IDF binary table. Tokens were rank-ordered

using the TF-IDF weight. Bigram detection was run using Matlab with ten significant bigrams added to the final corpus. Anchors extracted from the final free text dataset of clinical notes was done following the same natural language processing procedure. The top 25 anchors in each of these three categories were compared in their ability to estimate a multinomial model fit using the reported 500 randomized samples.

Chapter IV

RESULTS

Methods were modified to shift the focus from learning entire phenotype models for machine-based learning to the identification of free text anchors within a dataset comprised of physical therapy free text sources. This shift was made based on the sparsity of the dataset. This was confirmed by the observation that there were randomization models that did not run to completion. It was determined that this was due to the anchor token not being present within the patient's free text corpus. By removing such null token examples, all models were able to ultimately generate a thousand random experiment samples in multiple attempts further decreasing bias in the sample distribution. This practice conforms to the standards which were used by Halpern in the first two phases of his work.

The interpretation of the Predictive model results attempted to avoid violating the "winner takes all" principle. This principle of giving the algorithm the ability to choose only a single best value from a model's output goes against the spirit of identifying anchor terms. The ability of a range of anchors which are needed to identify a particular patient phenotype is the cornerstone of the anchor and learn framework. Choosing only

a single best value avoids the others every time. See Table 11 for a review of the maximum anchor value rankings. Since we expect a combination of winners when building a phenotype, a top range of values was needed for reporting purposes. Operationally, the top 20% of frequency values from 500 randomized samples were selected to be used in model reporting. This equated to the top five

TABLE 11. Anchor Token Maximum Frequency Analysis Rank Order			
Table	**Anchor**	**Rank**	**Percent of Total**
Expert Physical Therapist	filter	1	19.1
	develop	2	15.7
	cabg	3	11.2
	cad	4	10.1
	pulmonary	5	10.1
Academy of Acute Care Physical Therapy	dvt	1	29.8
	calf	2	19.0
	cancer	3	8.9
	prophylaxis	4	8.9
	compression	5	6.5
Clinician Documentation	filter	1	51.3
	leg	2	7.3
	ulcer	3	6.2
	wc_wheelchair	4	4.4
	foot	5	4.4
Ranked value count across 500 random samples in each table			

most frequently seen anchor variables out of the total twenty-five selections. Thereby, reporting of all anchor variable ranks will identify both the single "winner takes all" highest ranked variable and the ranking of the variables most often appearing in the top five. This will be the approach taken in the remainder of this chapter.

Expert Selected Anchors

The identified anchor terms derived from physical therapy free text data and ranked by a clinical expert did not match those reported in the literature. The top 20% most

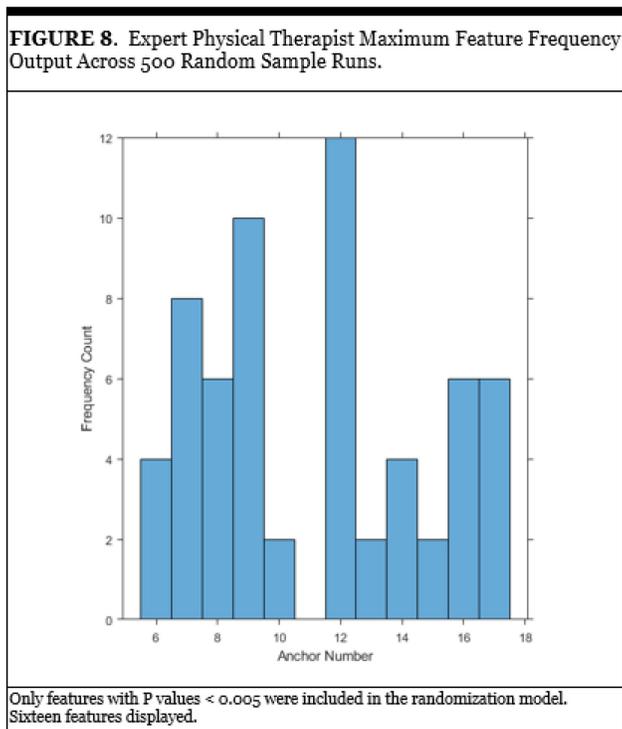| TABLE 12. Anchor Tokens Physician v. Physical Therapist			
Physician	**Feature Rank**	**Physical Therapist**	**Feature Rank**
dvt	1	boot	1
leg	2	movement	2
lovenox	3	develop	3
swelling	4	cad	4
calf	5	pulmonary	5
Abbreviation Key dvt = deep vein thrombosis cad = corinary artery disease			

TABLE 13. Anchors : Expert Selected					
Anchor	Log Coefficient	P value	Standard Error	95% Confidence Interval	
activity_tolerance	(0.47)	0.210	0.37	(1.20)	0.26
adl	(0.39)	0.703	1.02	(2.39)	1.61
ambulation	0.08	0.810	0.32	(0.55)	0.70
arthroplasty	(27.84)	0.932	326.90	(668.55)	612.88
bed_mobility	0.12	0.708	0.32	(0.51)	0.74
boot	0.87	0.200	0.68	(0.46)	2.21
cabg	(0.44)	0.683	1.08	(2.55)	1.67
cad	(1.19)	0.065	0.64	(2.45)	0.07
clot	(20.15)	0.068	11.06	(41.82)	1.52
co_sob	(32.03)	0.901	257.15	(536.04)	471.99
deep_vessels	9.61	1.000	189,812,531	(372,032,551)	372,032,570
develop	(0.52)	0.590	0.96	(2.41)	1.37
dvt	0.22	0.688	0.54	(0.84)	1.28
edema	0.14	0.762	0.46	(0.77)	1.05
filter	(0.05)	0.967	1.13	(2.26)	2.16
hypercoagul	2.17	0.703	5.68	(8.96)	13.30
lobe	(0.36)	0.731	1.05	(2.41)	1.69
movement	0.48	0.460	0.65	(0.80)	1.77
pe*	1.17	0.042	0.58	0.04	2.31
pulmonary	0.21	0.762	0.68	(1.12)	1.53
shin	(0.05)	0.964	1.10	(2.21)	2.11
thrombosis	(14.44)	0.971	402.36	(803.06)	774.18
vein	(14.92)	0.963	323.31	(648.61)	618.77
venous_embolism	(155.96)	1.000	189,812,531	(372,032,717)	372,032,405
vessel***	(20.99)	< .001	3.31	(27.48)	(14.50)
*P < 0.05 **P < 0.01 ***P < 0.001					

frequently observed anchor terms found in a prediction model for the VTE phenotype were found to have no overlap with the anchor tokens derived from physician free text data and can be viewed in Table 12. The expert selected sample of the top 25 anchors were compared in their ability to estimate a multinomial logistic regression model fit using one thousand randomized samples using 30 negative and 30 positive examples of records with VTE. The Log coefficient, P value, standard error, and 95% confidence intervals were reported for each of the 25 anchors and are reported in Table 13. Two of the anchor terms showed a statistically significant correlation with the presence of VTE: 'vessel' (P<0 .001) and 'pe' (P< 0.05).

FIGURE 8. Expert Physical Therapist Maximum Feature Frequency Output Across 500 Random Sample Runs.

Only features with P values < 0.005 were included in the randomization model. Sixteen features displayed.

Maximum frequency counts examined five hundred randomization runs where 16 valid terms were identified out of 25 which did not contain null value scores. The total frequency of appearance in the top 5 of all anchor terms was then found. Figure 8 shows the output. The most common anchor term was 'filter' (19.1%). Details can be seen in Table 11. The top 5 most frequent appearing anchor terms in descending order of total observed frequency were 'boot' (12.2%), 'movement' (10.4%), 'develop' (10.1%), 'cad' (9.5%) and 'pulmonary' (9.2%). See Appendix H for a complete list.

Academy of Acute Care Physical Therapy CPG Anchors

TABLE 14. Anchor Tokens Physician v. Academy of Acute Care Physical Therapy Clinical Practice Guideline.

Physician	Feature Rank	Clinical Practice Guideline	Feature Rank
dvt	1	dvt	1
leg	2	prophylaxis	2
lovenox	3	calf	3
swelling	4	compression	4
calf	5	swell	5

Abbreviation Key
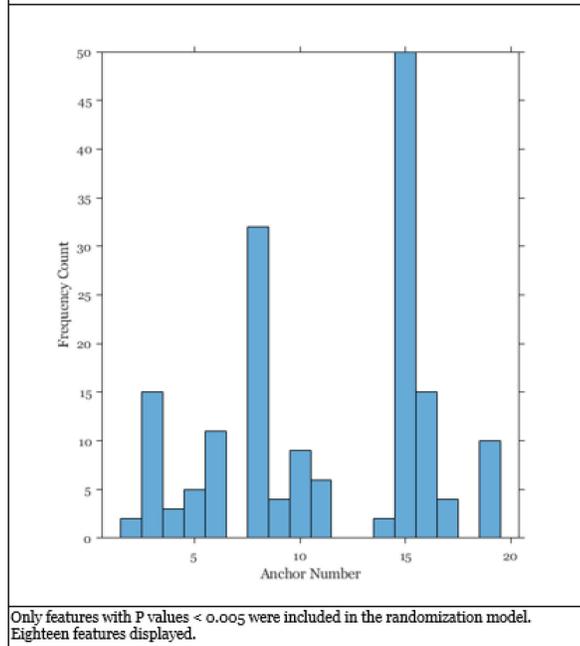dvt = deep vein thrombosis

The identified anchor terms derived from the association of acute care physical therapy clinical practice guidelines differed from those reported in the literature. See Table 14 for details. The top 20% most frequently observed anchor terms found in a prediction model for the VTE phenotype were found to have overlap with the anchor tokens derived from physician free-text data.

The Academy of Acute Care Physical Therapy Clinical Practice Guideline selected sample of the top 25 anchors were compared in their ability to estimate a multinomial logistic regression model fit using one thousand randomized samples using 30 negative and 30 positive examples of records with VTE. The Log coefficient, p-value, standard

TABLE 15. Anchors : Academy of Acutecare Physical Therapy CPG Selected.					
Anchor	Log Coefficient	P value	Standard Error	95% Confidence Interval	
ambulate*	(1.27)	0.034	0.60	(2.45)	(0.10)
anticoagul	(0.87)	0.359	0.95	(2.74)	0.99
calf	(0.40)	0.830	1.84	(4.00)	3.21
cancer	(0.41)	0.644	0.88	(2.13)	1.31
compression	(1.12)	0.305	1.10	(3.27)	1.02
dvt***	4.00	<0.001	0.43	3.16	4.83
edema*	(1.53)	0.013	0.62	(2.74)	(0.32)
embolus	(28.19)	0.998	14,282.83	(28,022.53)	27,966.15
extremity	(29.67)	0.997	7,340.18	(14,416.43)	14,357.09
fall**	(1.20)	0.002	0.40	(1.97)	(0.42)
family*	(1.55)	0.018	0.66	(2.83)	(0.26)
heparin	(34.64)	0.998	16,704.90	(32,776.26)	32,706.97
hospital	(0.58)	0.159	0.41	(1.39)	0.23
immobilization	(15.03)	1.000	24,845.31	(48,711.84)	48,681.78
inflammatory	(36.06)	0.995	6,326.22	(12,435.45)	12,363.33
leg*	0.94	0.026	0.42	0.11	1.76
mechanical**	2.24	0.008	0.84	0.59	3.89
pressure	0.79	0.189	0.60	(0.39)	1.97
prophylaxis	0.04	0.979	1.34	(2.59)	2.66
pulmonary	0.27	0.647	0.59	(0.88)	1.42
surgery	(1.00)	0.089	0.59	(2.16)	0.15
swell	1.24	0.059	0.66	(0.05)	2.53
ultrasound	(25.86)	0.994	3,293.67	(6,481.45)	6,429.74
vein	0.30	0.745	0.92	(1.50)	2.10
wells	(39.80)	0.995	6,581.19	(12,938.92)	12,859.32
* P < 0.05 ** P<0.01 *** P<0.001					

FIGURE 9. Academy of Acute Care Physical Therapy CPG Maximum Feature Frequency Output Across 500 Random Sample Runs.

Only features with P values < 0.005 were included in the randomization model. Eighteen features displayed.

error, and 95% confidence intervals were reported for each of the 25 anchors and can be seen in Table 15. Seven of the anchor terms showed statistically significant correlation with the presence of VTE : 'dvt' ($P<0.001$), 'fall' ($P<0.01$), 'mechanical' ($P<0.01$), 'edema' ($P<0.05$), 'family' ($P<0.05$), 'leg' ($P<0.05$), and 'ambulate' ($P<0.05$).

Maximum frequency counts examining five hundred randomization runs where 18 valid terms out of 25 did not contain null value examples. The total frequency of appearance in the top 5 of all anchor terms was found. Figure 9 reviews the frequency output. The most common anchor term was 'dvt' (29.8%). See Table 11 for details. The top 5 most frequent appearing anchor terms in descending order of total observed frequency were: 'dvt' (15.7%), 'prophylaxis' (10.3%), 'calf' (8.9%), 'compression' (6.9%), and 'swell' (6.7%). See Appendix H for a complete list.

Clinician Anchors

TABLE 16. Anchor Tokens Physician v. Clinician Documentation			
Physician	Feature Rank	Clinician	Feature Rank
dvt	1	filter	1
leg	2	ulcer	2
lovenox	3	weakness	3
swelling	4	endurance	4
calf	5	wc_wheelchair	5
Abbreviation Key dvt = deep vein thrombosis			

The identified anchor terms derived from the clinical free-text of the physical therapists differed from those reported in the literature and are detailed in Table 16. The top 20% most frequently observed anchor terms found in a prediction model for the VTE phenotype were found to have no overlap with the anchor tokens derived from physician free-text data.

The clinician documentation selected sample of the top 25 anchors were compared in their ability to estimate a multinomial

FIGURE 10. Clinician Maximum Feature Frequency Output Across 500 Random Sample Runs.

Only features with P values < 0.005 were included in the randomization model. Twenty-five features displayed.

logistic regression model fit using one thousand randomized samples using 30 negative and 30 positive examples of records with VTE. The Log coefficient, p-value, standard error, and 95% confidence intervals were reported for each of the 25 anchors are detailed
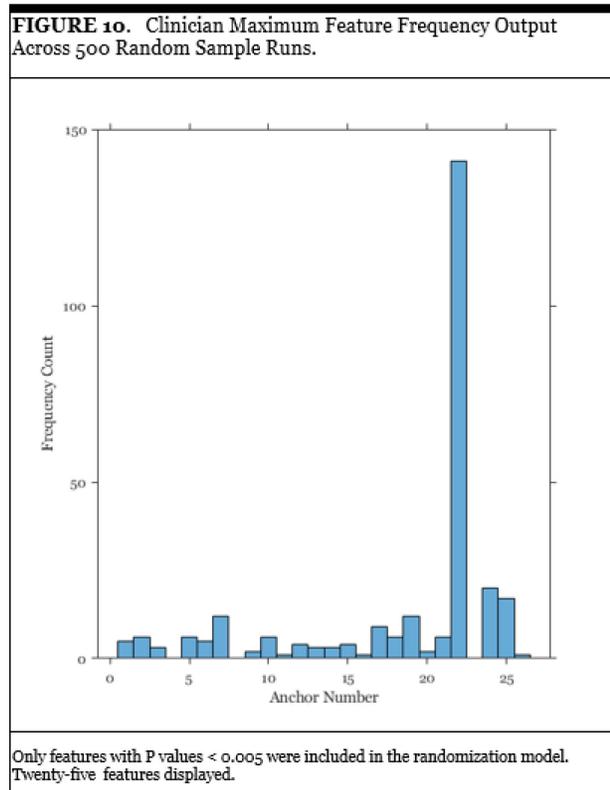
in Table 17. Four of the anchor terms showed a statistically significant correlation with

the presence of VTE: 'rom' (P<0.001), 'home' (P<0.05), 'le' (P<0.05), and 'sit' (P<0.05).

Maximum frequency counts examined five hundred randomization runs where the 25

valid anchor terms were most frequently seen. The total frequency of appearance in the

top 5 of all anchor terms was found. Figure 10 shows the output. The most common

term was 'filter'(51.3%), as can be seen in Table 11. The top 5 ranking anchor terms in

descending order of total observed frequency were: 'filter'(15.9%), 'ulcer'(6.2%),

'weakness'(5.6%), 'endurance'(5.5%), and 'wc_wheelchair'(4.9%). See Appendix H for a

complete list.

Anchor	Log Coefficient	P value	Standard Error	95% Confidence Interval	
TABLE 17. Anchors : Term Frequency – Inverse Document Frequency Selected.					
activity_tolerance	0.46	0.193	0.3518	(0.23)	1.15
acute	0.82	0.055	0.4273	(0.02)	1.66
ambulation	(0.69)	0.074	0.3891	(1.46)	0.07
anemia	0.58	0.139	0.3897	(0.19)	1.34
bed_mobility	(0.42)	0.369	0.4679	(1.34)	0.50
difficulty	(0.30)	0.419	0.3723	(1.03)	0.43
endurance	(0.29)	0.485	0.4190	(1.11)	0.53
fall	(0.24)	0.583	0.4282	(1.07)	0.60
fall_risk	(0.29)	0.500	0.4243	(1.12)	0.54
filter	(0.16)	0.731	0.4555	(1.05)	0.74
foot	0.23	0.552	0.3924	(0.54)	1.00
functional_activity	(0.55)	0.163	0.3963	(1.33)	0.22
home*	0.87	0.031	0.4025	0.08	1.66
hospitalize	(0.23)	0.554	0.3898	(0.99)	0.53
le*	0.91	0.040	0.4425	0.05	1.78
leg	(0.10)	0.816	0.4155	(0.91)	0.72
lle	(0.54)	0.147	0.3724	(1.27)	0.19
mobility	(0.02)	0.955	0.3512	(0.71)	0.67
pain	(0.10)	0.790	0.3849	(0.86)	0.65
rom***	3.06	< 0.001	0.5447	2.00	4.13
sit*	0.76	0.048	0.3850	0.01	1.52
time	(0.63)	0.106	0.3921	(1.40)	0.13
ulcer	(0.07)	0.889	0.4708	(0.99)	0.86
wc_wheelchair	0.74	0.056	0.3871	(0.02)	1.50
weakness	0.49	0.211	0.3910	(0.28)	1.26
*P < 0.05 **P<0.01 ***P<0.001					

Chapter V

DISCUSSION

Key Findings Summary

This study demonstrates that the methods used to perform anchor and learn phenotype identification, as published in the literature, are transferable to another electronic medical record within an alternate healthcare facility. The results of this study do not agree with previous research based on anchor term findings. The data indicate that the discovered physical therapy derived phenotype definition for venous thromboembolism (VTE) anchors did not mirror the existing physician derived phenotypes for VTE. The newly acquired anchor observation terms were statistically different when physical therapy free-text clinical notes were utilized for computable phenotype identification versus those obtained using physician free-text clinical notes.

Interpretation

This research provides a new insight into the relationship between anchor variables and the free text of physical therapists from which they are constructed. In line with the hypothesis, the top 5 maximum frequency analysis data shows that the cardiopulmonary

clinical expert physical therapist was concerned primarily with anchor variables, which represented the identification of patient signs and symptoms, decreased movement, and the identification of patient risk. These results mimic the existing recommendations put forth in clinical prediction guidelines.[125] In contrast, the anchor term, which was ranked first based on frequency by the expert, was 'filter.' If this result were the only one reported as would be the case in a "winner takes all" scenario, the expert would have been emphasizing a retrospective term that is significant for identification of a past medical history of VTE and surgical intervention.[32,33] If used as the primary basis for determination within a phenotypical model, it would not identify significant risk factors of developing VTE disease such as the immobilization or compression of a limb ('boot') and the significance of the level of mobility ('movement').[115,116] This result confirms the hypothesis that such anchor variables do not overlap with the physician experts.

In order to allow for validation of expert-selected anchor terms given the methodological constraints, two additional comparisons were necessary for clinical context. Based on the positive examples of it's use in the literature, the analysis was expanded to include natural language processing of the free text within the clinical practice guideline (CPG) itself, and on the entire corpus of clinician derived free text.[102,110,119] The top 5 maximum frequency analysis data shows that the free text found within the CPG was also concerned with anchor variables, which represented the identification of patient signs and symptoms and the identification of patient risk. As one would expect, the anchor term which was ranked first based on frequency by the CPG was 'dvt.' These found anchor terms had significant overlap with the physician selected anchors. The real value of the CPG maximum frequency analysis did not lie in the 3 out of 5 terms 60% match with the physician terms, but with the lack of variable overlap with

the expert physical therapy clinician. It should be reasonably assumed that an expert in cardiopulmonary physical therapy would have embedded the values and terminology reflected by published evidence-based practice guidelines and selected similar terms. Again, this was not the case, and no matches were found between the expert physical therapist, and CPG discovered anchor terms.

The answer as to why the physical therapy expert anchor variables did not overlap with the physician anchors and why the found anchors within the CPG did match, can be found in the analysis of the entire corpus of clinician derived free text. Recall that the top 5 maximum frequency analysis data shows that the clinician documentation pointed toward an identification of the presence and sequela of decreased movement and past medical history for VTE. Clinicians tended to focus on history and treatment for lack of movement leading to a lower extremity origin of VTE. As movement specialists, physical therapists are trained to view the patient within this context.[13] The clinician derived free text was discovered to have no matches with the expert physical therapist and the physician based anchor terms. This lends evidence towards the assertion that clinicians are not aware of the CPG either through lack of knowledge or due to a general awareness that the CPG itself existed in the literature.[14, 16, 120] Despite the reasoning for the lack of overlap between the clinician based terms with the CPG, it is evidence of the identified need for such clinical decision support tools at the bedside.[16]

Discussion

The results of this work met the expectations of this researcher. Physical therapists and physicians are trained to place the patient at the center of care. Each discipline views the patient with a distinct set of professional tools and language for documenting their

treatments. The practice disparities between professions are converging and are closer together than ever. However, within the discipline of biomedical informatics, significant differences do exist as the specificity of tools improves, as confirmed by this research. As this precision grows and modeling advances, so must the precision of the variables used to create such models.

There were vital anchor tokens that did not appear in each of the corpora analyzed. The presence of tokens representing features existing within the Wells rule definition was noticeably absent. Edema related anchor terms were only identified in clinical notes and not by the expert physical therapist and CPG anchor terms. A hallmark sign of LE DVT origin of VTE is pitting edema.[27, 28] The stem 'pit' is noticeably absent in the top 25, statistically ranked values in each of the datasets examined. Terms relating to pressure garments, stockings, and TED Hose were sparse. Patient education and interprofessional communication-related terms were also strangely absent in the top 25 statistically found anchor tokens. The absence of these terms is suspect given the function of the treatment setting within a long term care facility. However, it very well could be attributed to the small size of the physical therapy department and the limited diversity of free text examples of staff documentation. Another unidentified contributor to a lack of terms appearing may be due to the configuration of the EMR itself. Clinicians may have documented these findings in another place, within an embedded form, in another section of the patient's chart, or in a method of drop-down or checklist box type documentation which did not lend itself to free text status.

Clinical decision support tools are vital throughout all treatment settings to help clinicians make evidence-based decisions at the bedside. Hospitals and emergency rooms appear initially to be an excellent place to start to define phenotypes. Acute situations deserve research attention to achieve optimal outcomes. However, based on

106

the findings of this research, a more plausible shift in focus for this work lies within the early identification of disease before the need arises for acute care in the first place.

These results should be taken into account when considering the performance of such tools at the bedside. Ultimately, there was no overlap in the top 20% of terms between physical therapists and physicians. Clinical decision support in a real-world clinical setting would, therefore, not have alerted the physical therapy clinician to the presence of VTE if embedded within the EMR using the physician anchor terms. This represents a real-world, life or death consequence of the inability of current systems to adapt to alternate practice settings and professionals who use them. The urgent need exists to acquire the ability to predict the presence of a disease state early, which may lead to optimal patient outcomes, decreased morbidity and prolonged quality of life.[89]

Implications

VTE is a life-threatening condition with an incidence of mortality within one month of diagnosis of 10%-30%.[126] Half of all patients diagnosed with a LE DVT origin of VTE have life long complications. Approximately 33% of patients will experience another VTE within ten years.[127]

Physical therapists encounter patients who are at risk for or have a history of VTE in all institutional settings and across all specialty areas of practice. Physical therapy treatment revolves around patient movement, and this is the routine focus to both prevent and facilitate recovery from the diagnosis of a VTE.

According to the clinical practice guidelines put forth by the academy of acute care physical therapy, the physical therapist plays a significant role in identifying patients who are at high risk for a VTE. Specific treatment algorithms exist which are ripe for

integration within a clinical decision support tool creation pathway and are presented in Appendix I. Evidence-based practice through the use of the recommended guidelines within the CPG dictate that physical therapists should be aware of the signs and symptoms of a LE DVT. When signs and symptoms are present, the likelihood of a LE DVT should be determined through the Wells criteria for LE DVT. Results should be shared with the interprofessional team to consider treatment options.[125]

In patients with a diagnosed LE DVT, once a medication's therapeutic levels or an acceptable period is reached after administration, mobilization should begin. Although there are risks associated with mobilization, the risk of inactivity is higher. Complications following LE DVT can continue for years or even a lifetime. Physical therapists can help decrease these complications through education, mechanical compression, and exercise.

Before this research, there did not exist an established set of anchors with a likelihood of detecting the presence of LE DVT originated VTE accurately within the profession of physical therapy. An initial listing of such anchor variables has now been discovered. Further analysis is needed to expand and document the ability of machine learning classifiers to learn predictive models to identify those patients at risk and who have active VTE disease.

Limitations

The methodological choices were constrained not by the ability to gain access to years worth of data, but ultimately by the sparsity of the dataset. Subsequently, the methods were adjusted to allow for random sampling to reduce the issue of overfitting. This is a real-world problem that is encountered frequently by data scientists.[74] Best practice,

when facing sparse datasets, is to eliminate randomization models that do not meet a predetermined significance level.[128] In this case, randomized cases that yielded statistical models with P values of <0.005 were excluded in the run count. By controlling the randomization in this manner, if anchor variables ultimately appeared in a randomized model, they were significant. Multiple appearances of these anchor variables in the top 20% of terms are confirmation of their continued significant presence.[87] Feature weights reported in the literature signify the relative importance of that term in a statistical model. Frequency counts within this methodology yielded excellent comparison results for feature identification and comparison.

It was assumed that a large database was used by Halpern to learn anchors. This, in fact, was not the case. Halpern reported that depending upon the phenotype, his test sets ranged in size from 1,082 to 62,589 patients, having a variable number of patients available for training. Where Halpern did have an advantage was with having access to a 200,000 patient encounter database from which to train his machine learning classifier. In this research adjustments in methods were made to accommodate a smaller test set of 1,668 patients. The randomization model yielded a competitive sample of 500 (90,000 patient encounters – 60 samples x 500 random runs x 3 datasets) was used across all three data types. The "winner takes all" top 25 terms statistical analysis utilized 1,000 random samples. (180,000 patient encounters – 60 samples x 1000 random runs x 3 datasets)"

It was beyond the scope of this study to utilize the computerized anchor elicitation tool, as reported in the literature for real-time expert feedback of model analysis. The utility of this tool was to allow for the quick survey of a random sample of individual patient records and associated predictions based on a modifiable set of anchor features. The methodological choice to replace this step with a commonly used survey tool (Survey

Monkey) was in line with the original intent of the published research. To confirm that the instrument was yielding the intended results, interrater reliability with an alternate expert clinician demonstrated good reliability of term selection (κ =0.8430). The results show that a commonly used tool was able to select anchor terms in an acceptable alternate manner. It was not able to show results in real-time to the expert physical therapist which may have increased the accuracy of the anchor term selection.

The concatenation of a combined corpus size of 650,000 tokens from two disparate databases proved to be technically and administratively burdensome. The proprietary nature of databases delivered within a software as a service arrangement was a significant barrier. This was in spite of the meaningful use of electronic medical records for clinical documentation by a technically savvy facility. A lack of administrative level access to data, which was in a proprietary relational database format necessitated extensive negotiation with the vendor for access. Ultimately the widespread use of data manipulation before pre-processing became an inevitable step in the data preparation. These types of data governance and ownership issues are not always transparent to a healthcare provider and pose a significant barrier to the availability of useful tools for future data analysis methods.

The utilization of natural language processing methods is integral to the success of the future of patient phenotyping.[89, 102, 103, 110, 119, 122] The methods sections described by Halpern et al. in the literature were void of the specificity needed to reproduce the experiments with certainty. Caution should be used when deploying standard stop word filters such as the Natural Language Toolkit, Standford Part of Speech Tagger, or Porter stemmer in base form. [59, 129-131] Unmodified stop word filters trim indiscriminately when applied to clinical free-text data pre-processing. These tools have great utility in the elimination of single letters, which are left as a by-product of word stemming and

lemmatization. Removal of such individual letters such as 'a,' 'd,' 'i,' 'o,' 's' and single orphaned terms such as 'don' are familiar to each tool. These are discipline-specific abbreviations within the physical therapy profession for terms such as assistance, dependant, independent, oriented, supervision and the ability to don clothing. Removal of these terms can strip significant clinical context from free-text datasets and significantly impact the definition of anchor terms used in phenotype discovery.

Recommendations

Future research using the anchor and learn method needs to take heed of several learned examples found in this work. Interoperability is needed for small organizations that have added tools sequentially over time. The opportunity exists for software vendors to address this issue and thereby improve the delivery of evidence-based practice. The anchor and learn framework ultimately requires a large dataset in order to allow for machine learning predictive models. Frequency counts, as demonstrated by this work, can be quickly determined as an initial phase of data analysis before training learned models. Before learning feature weights, finding feature frequencies is a quick and efficient method for new dataset analysis. Taking the top 20% of frequently seen anchor terms produced markedly different results than using a "winner takes all" approach. Given the significant appearance of movement-based anchor terms, future work within the physical therapy field using more extensive datasets that will allow for machine learning to take place is supported.

Chapter VI

SUMMARY AND CONCLUSIONS

The delivery of patient-centered care requires an interdisciplinary team of clinicians to successfully achieve optimal patient outcomes. Evidenced-based practice is enhanced by the presence of clinical decision support tools in the clinical workflow of the modern healthcare system. This work identifies that a real need exists to identify anchor terms to allow for the identification of patients within a given phenotype using the full power and benefit of modern machine learning techniques. It is necessary to expand the definitions of phenotypical anchor terms to allow for shared evidence-based practice tool use across disciplines to achieve optimal patient care.

A call for further detail in methods by published works must be adopted by informatics researchers to allow for duplication and adoption of methods. The black box of machine learning algorithms must become opaque quickly.

REFERENCES

1. HealthIT.gov. Meaningful Use and the Shift to the Merit-based Incentive Payment System. https://www.healthit.gov/topic/federal-incentive-programs/meaningful-use. Accessed December 2018.

2. eHealthUniversity. Centers for Medicare & Medicaid Services. An Introduction to Medicare EHR Incentive Program for Eligible Professionals. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/EHR_Medicare_Stg1_BegGuide.pdf. Accessed January 2019.

3. Charles D, Gabriel M, Furukawa MF. Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013. ONC Databrief:16.May 2014. https://www.healthit.gov/sites/default/files/briefs/oncdatabrief16.pdf. Accessed January 2019.

4. Comstock J. MobiHealthNews. With $4.5B, Q3 2019 was digital health funding's biggest quarter yet. October 11, 2018. https://www.mobihealthnews.com/content/45b-q3-2018-was-digital-health-fundings-biggest-quarter-yet. Accessed January 2019.

5. National Healthcare Expenditure Projections, 2017-2026. Centers for Medicare and Medicaid Services, Office of the Actuary. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsProjected.html. Accessed January 2019.

6. HealthIT.gov. MACRA: What is the MACRA Quality Payment Program. https://www.healthit.gov/topic/federal-incentive-programs/macra. Accessed January 2019.

7. Quality Payment Program. Centers for Medicare & Medicaid Services. QPP Key Objectives Information Sheet. https://qpp.cms.gov/about/qpp-overview. Accessed January 2019.

8. Coleman K, Wagner E, Schaefer J, Reid R, LeRoy L. Redefining Primary Care for the 21 st Century. White Paper. (Prepared by Abt Associates, in partnership with the MacColl Center for Health Care Innovation and Bailit Health Purchasing, Cambridge, MA under Contract No.290-2010-00004-I/ 290-32009-T.) AHRQ Publication No. 16(17)-0022-EF. Rockville, MD: Agency for Healthcare Research and Quality; October 2016.

9. UHealth. The University of Utah. Value Equation. https://uofuhealth.utah.edu/value/value-equation.php. Accessed January 2019.

10. Porter ME. What is value in health care? N Engl J Med. 2010;363(26):2477-81.

11. The White House. Office of the Press Secretary. FACT SHEET: Obama Administration Announces Key Actions to Accelerate the Precision Medicine Initiative. https://obamawhitehouse.archives.gov/the-press-office/2016/02/25/fact-sheet-obama-administration-announces-key-actions-accelerate. Accessed January 2019.

12. APTA. http://www.apta.org/Default.aspx. Accessed August 12, 2016.

13. Bellamy J. Vision Statement for the Physical Therapy Profession and Guiding Principles to Achieve the Vision. http://www.apta.org/Vision/. Accessed August 12, 2016.

14. Green S, McDonald S, Holland AE, Xs M. Informing physiotherapy decisions with reliable evidence: how physiotherapists have contributed to Cochrane and how Cochrane has informed evidence-based physiotherapy. *J Physiother*. 2014;60(1):1-4.

15. Gardner K. American Board of Physical Therapy Specialties - ABPTS. http://www.abpts.org/home.aspx. Accessed August 12, 2016.

16. Knox GM, Snodgrass SJ, Rivett DA. Physiotherapy clinical educators' perceptions and experiences of clinical prediction rules. *Physiotherapy*. 2015;101(4):364-372 9p. doi:10.1016/j.physio.2015.03.001.

17. Stiell IG, McKnight R, Greenberg GH, et al. Implementation of the Ottawa ankle rules. *JAMA*. 1994;271(11):827-832.

18. Wells PS, Hirsh A, Anderson DR, et al. A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process. *J Intern Med*. 1998;243:15-23.

19. Stiell IG, Wells GA, Hoag RH, et al. Implementation of the Ottawa knee rule for the use of radiography in acute knee injuries. *JAMA*. 1997;278(23):2075-2079.

20. Bernhardsson S, Larsson MEH, Eggertsen R, et al. Evaluation of a tailored, multi-component intervention for implementation of evidence-based clinical practice guidelines in primary care physical therapy: a non-randomized controlled trial. *BMC Health Serv Res*. 2014;14:105.

21. Halpern Y, Choi Y, Horng S, Sontag D. Using Anchors to Estimate Clinical State without Labeled Data. AMIA Annu Symp Proc. 2014 November 14, 2014:606-15.

22. Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *The Lancet*. // 1997;350(9094):1795-1798.

23. Kane D, Balint PV, Gibney R, Bresnihan B, Sturrock RD. Differential diagnosis of calf pain with musculoskeletal ultrasound imaging. *Annals of the Rheumatic Diseases*. January 1, 2004 2004;63(1):11-14.

24. Halpern Y, Horng S, Choi Y, Sontag D. EMR Phenotyping using Anchor & Learn Framework Halpern, Appendix1, Methodology Details. https://academic-oup-com.proxy.libraries.rutgers.edu/jamia/article/23/4/731/2200279. Accessed June 2019.

25. Agency for Healthcare Research and Quality. Patient-Centered Decision Support. https://pccds-ln.org/node/276. Accessed June 2019.

26. National Institutes for Health. National Center for Complementary and Integrative Health. Clinical Practice Guidelines. https://nccih.nih.gov/health/providers/clinicalpractice.htm. Accessed June 2019.

27. American Academy of Family Physicians. Clinical Practice Guideline. Diagnosis of Venous Thromboembolism- Clinical Practice Guideline. https://www.aafp.org/patient-care/clinical-recommendations/all/venous-thromboembolism1.html. Accessed June 2019.

28. American Academy of Family Physicians. Clinical Practice Guidelines. https://www.aafp.org/patient-care/browse/type.tag-clinical-practice-guidelines.html. Accessed June 2019.

29. Childs JD, Cleland JA, Development and Application of Clinical Prediction Rules to Improve Decision Making in Physical Therapist Practice, Physical Therapy, Volume 86, Issue 1, 1 January 2006, Pages 122–131, https://doi-org.proxy.libraries.rutgers.edu/10.1093/ptj/86.1.122

30. Richesson R, Smerek M. Duke University. Rethinking Clinical Trials. Electronic Health Records-Based Phenotyping. https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/. Accessed June 2019.

31. Office of the Surgeon General (US); National Heart, Lung, and Blood Institute (US). The Surgeon General's Call to Action to Prevent Deep Vein Thrombosis and Pulmonary Embolism. Rockville (MD): Office of the Surgeon General (US); 2008. Available from: https://www-ncbi-nlm-nih-gov.mt.opal-libraries.org/books/NBK44178/

32. Piazza G, Goldhaber SZ. Acute pulmonary embolism: part I: epidemiology and diagnosis.Circulation. 2006;114(2):e28–32.

33. Piazza G, Goldhaber SZ. Acute pulmonary embolism: part II: treatment and prophylaxis.Circulation. 2006;114(3):e42–7.

34. Mohr DN, Silverstein MD, Heit JA, Petterson TM, O'Fallon WM, Melton LJ. The venous stasis syndrome after deep venous thrombosis or pulmonary embolism: a population-based study. Mayo Clin Proc. 2000;75(12):1249–56. (3)

35. Prandoni P, Lensing AW, Cogo A, Cuppini S, Villalta S, Carta M, et al. The long-term clinical course of acute deep venous thrombosis. Ann Intern Med. 1996;125(1):1–7. (4)

36. Heit JA, Mohr DN, Silverstein MD, Petterson TM, O'Fallon WM, Melton LJ 3rd. Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study. Arch Intern Med. 2000;160(6):761–8. (5)

37. Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Predictors of survival after deep vein thrombosis and pulmonary embolism: a population-based, cohort study. Arch Intern Med. 1999;159(5):445–53. (6)

38. Agency for Healthcare Research and Quality. Effective Health Care Program. Chapter 12: Systematic Review of Prognostic Tests. https://effectivehealthcare.ahrq.gov/topics/methods-guidance-tests-prognostic/methods. Accessed June 2019.

39. Agency for Healthcare Research and Quality. Effective Health Care Program. Chapter 8: Meta-Analysis of Test Performance When There Is a "Gold Standard." https://effectivehealthcare.ahrq.gov/topics/methods-guidance-tests-metaanalysis/methods. Accessed June 2019.

40. EHR Intelligence. Xtelligent Healthcare Media. EMR v. EHR: Electronic Medical, Health Record Differences. https://ehrintelligence.com/features/emr-v.-ehr-electronic-medical-health-record-differences?eid=CXTEL000000340177&elqCampaignId=8658&elqTrackId=c9f5a45268e042ee9c2ed6c3677ffb28&elq=d1938c4abb0c49a18acef4c1bf0547f9&elqaid=9124&elqat=1&elqCampaignId=8658. Accessed June 2019.

41. Pocket Glossary of Health Information Management and Technology, Fifth Edition. AHIMA Press. June 23, 2017.

42. First Databank. http://www.fdbhealth.com. Accessed June 2019.

43. HIPAA Privacy Deidentification. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard. Accessed July 2019.

44. American Academy of Professional Coders. Translator Tool. https://www.aapc.com/icd-10/codes/. Accessed June 2019.

45. Hinz ERM, Bastarache L, Denny JC. A Natural Language Processing Algorithm to define a Venous Thromboembolism Phenotype. AMIA Annual Symposium Proceedings. 2013;2013:975.

46. World Health Organization. Classifications. Classification of Diseases. https://www.who.int/classifications/icd/en/. Accessed June 2019.

47. Centers for Disease Control and Prevention. National Center for Health Statistics. ICD. https://www.cdc.gov/nchs/icd/icd10cm.htm. Accessed June 2019.

48. Goldberg Y. Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Toronto, Canada. Morgan & Claypool Publishers, April 17, 2017.

49. Ticary Solutions, LLC. What is Natural Language Processing (NLP)? Blog post. https://ticary.com/2017/12/12/what-is-nlp.html. Accessed June 2019.

50. Google Code Archive. https://code.google.com/archive/p/word2vec/. Accessed June 2019.

51. Chen E. Introduction to Latent Dirichlet Allocation. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/. Accessed June 2019.

52. Algorithmia. Blog August 11, 2016. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/. Accessed June 2019.

53. Shetty B. Towards Data Science. Natural Language Processing (NLP) for Machine Learning. https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b. Accessed June 2019.

54. SAS Institute. SAS Insights: Analytics Insights. Natural Language Processing (loT) What it is and why it matters. https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html. Accessed June 2019.

55. YouTube. Quantopian. Bag-of-Words. https://youtu.be/IRKDrrzh4dE. Accessed June 2019.

56. FreeCodeCamp. An introduction to Bag of Words and how to code it in Python for NLP. December 18, 2018. https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/. Accessed June 2019.

57. Brownlee J. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/. Accessed June 2019.

58. Deng Y, Denecke K. Visualizing unstructured patient data for assessing diagnostic and therapeutic history. Stud Health Technol Inform. 2014;205:1158-62.

59. Stanford Part-Of-Speech Tagger. Stanford University. GitHub. https://stanfordnlp.github.io/CoreNLP/. Accessed June 2019

60. What does tf-idf mean? http://www.tfidf.com/. Accessed September 2019

61. Google Code Archive. NegEx. https://code.google.com/archive/p/negex/. Accessed June 2019.

62. Agency for Healthcare Research and Quality. Patient Safety Indicators Overview. Updated Patient Safety Indicators Technical Specifications. PSI 12 Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate. https://www.qualityindicators.ahrq.gov/Modules/PSI_TechSpec_ICD10_v2018.aspx. Accessed June 2019.

63. Porter ME. What is value in health care? N Engl J Med. 2010;363(26):2477-81.

64. Donaldson MS, Yordy KD, Lohr KN, et al. eds. Primary Care: America's Health in a New Era. Report of a study by a committee of the Institute of Medicine, Division of Health Care Services: National Academy Press; 1996.

65. Kearon C. Natural history of venous thromboembolism. Circulation. 2003;107(23 Suppl 1):I22–30.(7)

66. Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Predictors of survival after deep vein thrombosis and pulmonary embolism: a population-based, cohort study. Arch Intern Med. 1999;159(5):445–53.(6)

67. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al., The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genomics 4 (2011) 13. Epub 2011/01/29.

68. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al., The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future, Genet. Med.: Off. J. Am. College Med. Genet. 15 (10) (2013) 761–771. Epub 2013/06/08.

69. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J, Some experiences and opportunities for big data in translational research, Genet. Med.: Off. J. Am. College Med. Genet. 15 (10) (2013) 802–809. Epub 2013/09/07.

70. As accessed at https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE

71. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, et al., The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities, AMIA Annu. Symp. Proc./AMIA Symp. AMIA Symp. 2011 (2011) 248–256. Epub 2011/12/24.

72. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al., Normalization and standardization of electronic health records for high throughput phenotyping: the SHARPn consortium, J. Am. Med. Inform. Assoc:JAMIA 20 (e2) (2013) e341–e348. Epub 2013/11/06.

73. Thompson EE, Steiner JF, Embedded research to improve health: the 20th annual HMO Research Network conference, March 31–April 3, 2014, Phoenix, Arizona, Clin. Med. Res. 12 (1–2) (2014) 73–76. Epub 2014/10/30.

74. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook CM, Hart G, et al., The HMO research network virtual data warehouse: a public data model to support collaboration, EGEMS (Wash, DC) 2 (1) (2014) 1049. Epub 2014/01/01.

75. Daugherty SE, Wahba S, Fleurence R, Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research, J. Am. Med. Inform. Assoc.: JAMIA 21 (4) (2014) 583–586. Epub 2014/05/14.

76. Bambrick N. KDnuggets Newsletters. Support Vector Machines – What are They? https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html. Accessed June 2019.

77. Techopedia. Support Vector Machine (SVM). https://www.techopedia.com/definition/30364/support-vector-machine-svm. Accessed June 2019.

78. Patel S. Machine Learning 101. Chapter 2: SVM (Support Vector Machine) – Theory https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. Accessed June 2019.

79. Institute for Healthcare Improvement. Triple Aim for Populations. http://www.ihi.org/Topics/TripleAim/Pages/default.aspx. Accessed June 2019.

80. Whittington JW, Nolan K, Lewis N, Torres T. Pursuing the Triple Aim: The First 7 Years. The Milbank Quarterly. 2015;93(2):263-300 https://www-ncbi-nlm-nih-gov.proxy.libraries.rutgers.edu/pmc/articles/PMC4462878/

81. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health affairs. May-Jun 2008;27(3):759-769.

82. Coleman K, Wagner E, Schaefer J, Reid R, LeRoy L. Redefining Primary Care for the 21st Century. Nonnenhorn: TCP Terra-Consulting-Partners; 2016 October.

83. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. 2014;12(6), 573-6. doi: 10.1370/afm.1713.

84. Wagner EH. The Fourth Aim: Primary Care and the Future of American Medical Care. Dallas, TX: IHI Office Practice Summit; 2012.

85. American Heart Association. What is Venous Thromboembolism (VTE)? https://www.heart.org/en/health-topics/venous-thromboembolism/what-is-venous-thromboembolism-vte. Accessed June 2019.

86. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. ACM; Aug 24, 2008.

87. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney T, Gyang E, Shah NH. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016;23:116-1173.

88. Halpern Y, Choi Y, Horng S, Sontag D. Using Anchors to Estimate Clinical State without Labeled Data. AMIA Annu Symp Proc. 2014 November 14, 2014:606-15.

89. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python; 2011. http://scikit-learn.org. Accessed January 2019.

90. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014 Mar;21(2):221-30.

91. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS ONE. 2018;13(2):e0192360.

92. Lipton ZC. The mythos of model interpretability. https://arxiv.org/pdf/1606.03490.pdf. Accessed June 2019.

93. Goodman B, Flaxman S. EU regulations on algorithmic decision-making and a "right to explanation." In ICML Workshop on Human Interpretability in Machine Learning. (WHI 2016);2016. Available at: https://ui.adsabs.harvard.edu/abs/2016arXiv160608813G/abstract. Accessed June 2019.

94. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016 Nov;23(6):1046-52.

95. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Medical Informatics and Decision Making 2017;17-24.

96. Paparrizos J, White RW, Horvitz E. Screening for pancreatic Adenocarcinoma using signals from web search logs: feasibility study and results. J Oncol Pract. 2016;12:737-44.

97. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA. 2017 August 8, 318(6):517-8.

98. Povyakalo AA, Alberdi E, Strigini L, et al., How to discriminate between computer-aided and computer-hindered decisions. Med Decis Making 2013;33(1):98-107.

99. Tsai TL, Fridsma Db, Gatti G. Computer decision support as a source of interpretation error. J Am Med Inform Assoc. 2003;10(5)478-483.

100. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare:predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21sh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Cham, Switzerland: Springer International Publishing AG: 2015:1721-1730.

101. Esteban S, Rodriguez Tablado M, Ricci RI, Terrasa S, Kopitowski K. A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases. BMC Res Notes. 2017 July 14, 10(1):281.

102. Kukhareva P, Staes C, Noonan KW, Mueller HL, Warner P, Shields DE, et al. Single-reviewer electronic phenotyping validation in operational settings: Comparison of strategies and recommendations. J Biomed Inform. 2017;66:1-10.

103. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ : British Medical Journal. 2015 Apr 24,;350(apr24 11):h1885.

104. Alnazzawi N, Thompson P, Batista-Navarro R, Ananiadou S. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. From Louhi 2014: The Fifth International Workshop on Health Text Mining and Information Analysis Gothenburg, Sweden. 27 April 2014. In: BMC Medical Informatics and Decision Making 2015;15(2):S3.

105. Travers DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. J Biomed Inform. 2003;36(4-5):260-70.

106. Hinz ERM, Bastarache L, Denny JC. A Natural Language Processing Algorithm to define a Venous Thromboembolism Phenotype. AMIA Annual Symposium Proceedings. 2013;2013:975.

107. Kotfila C, Uzuner O. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. J Biomed Inform. 2015 December;58(Suppl):S102.

108. Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. Journal of Biomedical Informatics. 2016 Dec;64:168-78.

109. Lin C, Hsu C, Lou Y, Yeh S, Lee C, Su S, et al. Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. J Med Internet Res. 2017 November 6, 19(11):e380.

110. Mehrabi S, Schmidt CM, Waters JA, Beesley C, Krishnan A, Kesterson J, et al. An efficient pancreatic cyst identification methodology using natural language processing. Stud Health Technol Inform. 2013;192:822-6.

111. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying Patients with Depression Using Free-text Clinical Documents. Stud Health Technol Inform. 2015;216:629-33.

112. Kontio E, Airola A, Pahikkala T, Lundgren-Laine H, Junttila K, Korvenranta H, et al. Predicting patient acuity from electronic patient records. J Biomed Inform. 2014 October;51:35-40.

113. Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, Pacheco JA, et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. J Biomed Inform. 2016 Aug;62:232-42.

114. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017:48-57.

115. Mate S, Castellanos I, Ganslandt T, Prokosch H, Kraus S. Standards-Based Procedural Phenotyping: The Arden Syntax on i2b2. Stud Health Technol Inform. 2017;243:37-41.

116. Lecumberri R, Panizo E, Gomex-Guiu A, Varea S, Garcia-Quetglas E, Serrano M, et al. Economic impact of an electronic alert system to prevent venous thromboembolism in hospitalized patients. Journal of Thrombosis and Haemostasis. 2011 Jun;9(6):1108-15.

117. Umscheid CA, Hanish A, Chittams J, Weiner MG, Hecht TEH. Effectiveness of a novel and scalable clinical decision support intervention to improve venous thromboembolism prophylaxis: a quasi-experimental study. BMC medical informatics and decision making. 2012 Aug 31; 12(1):92.

118. As accessed at https://www.qualityindicators.ahrq.gov/Modules/PSI_TechSpec_ICD10_v201 8.aspx on 6/8/2019.

119. Kucher N, Puck M, Blaser J, Bucklar G, Eschmann E, Lüscher TF. Physician compliance with advanced electronic alerts for preventing venous thromboembolism among hospitalized medical patients. Journal of thrombosis and hemostasis: JTH. 2009 Aug;7(8):1291-6.

120. Goldbraich E, Waks Z, Farkash A, Monti M, Torresani M, Bertulli R, et al. Understanding Deviations from Clinical Practice Guidelines in Adult Soft Tissue Sarcoma. Stud Health Technol Inform. 2015;216:280-4.

121. Planquette B, Maurice D, Peron J, Mourin G, Ferre A, Sanchez O, et al. Knowledge of the diagnostic algorithm for pulmonary embolism in primary care. Eur J Intern Med. 2015 January;26(1):18-22.

122. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA Oct 20 1999; 282(15):1458-65.

123. North F, Fox S, Chaudhry R. Clinician time used for decision making: a best-case workflow study using cardiovascular risk assessments and Ask Mayo Expert algorithmic care process models. BMC Med Inf Decis Mak. 2016 July 20; 16:96.

124. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform, 2001 Oct;34(5):301-10.

125. Academy of Acute Care Physical Therapy Clinical Practice Guideline (CPR) for Venous Thromboembolism. (As Accessed on 1/4/19 at https://www.acutept.org/page/VTEGuidelines?)

126. Beckman MG, Hooper WC, Critchley SE, Ortel TL. Venous thromboembolism: a public health concern. Am. J. Prev. Med. 2010;38(4):S495-501.

127. Kahn SR, Ducruet T, Lamping DL, et al. Prospective evaluation of the health-related quality of life in patients with deep venous thrombosis. Arch. Intern. Med. 2005;165(10):1173- 1178.

128. Lee CH, Yoon HJ. Medical big data: promise and challenges. *Kidney Res Clin Pract*. 2017 Mar;36(1):3-11. doi: 10.23876/j.krcp.2017.36.1.3. Epub 2017 Mar 31.

129. Natural Language Toolkit as accessed at https://www.nltk.org/ on 9/30/2019.

130. Sparck Jones K, Willet P, *Readings in Information Retrieval*, San Francisco: Morgan Kaufmann,

131. Porter Stemmer as accessed at https://tartarus.org/martin/PorterStemmer/ on 9/30/2019.

APPENDIX

Appendix A

Electronic Medical Record Phenotyping using the Anchor & Learn Framework

Yoni Halpern, Steven Horng, Youngduck Choi, David Sontag
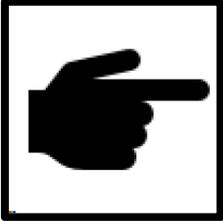
Phenotype Definitions

Phenotype	Data Source	Anchors
TABLE 18. Electronic Medical Record Phenotyping using the Anchor & Learn Framework | Phenotype Definitions.		
Phenotype	**Data Source**	**Anchors**
Abdominal pain		540-543:appendicitis
		560.0:intussusception
		560.2:volvulus of intestine
		560.89:intestinal obstruct nec
		560.9:intestinal obstruct nos
		562.01:diverticulitis s intest no hem
		562.03:diverticulitis sm intest w/hem
		562.11:diverticulitis colon-no hem
		562.13:diverticulitis colon w/hem
		574:cholelithiasis
		575.0:acute cholecystitis
		575.10:cholecystitis, unspecified
		576.1:cholangitis
		577.0:acute pancreatitis
		789.00:abdominal pain unspec site
		789.01:abdominal pain ruq
		789.02:abdominal pain luq
		789.03:abdominal pain rlq
		789.04:abdominal pain llq
		789.05:abdominal pain periumbilic
		789.06:abdominal pain epigastric
		789.07:abdominal pain generalized
		789.09:abdominal pain other specied

124

		789.0:abdominal pain
		789.60:abdominal tenderness unsp site
		789.61:abdominal tenderness ruq
		789.62:abdominal tenderness luq
		789.63:abdominal tenderness rlq
		789.64:abdominal tenderness llq
		789.65:abdominal tenderness periumbilic
		789.66:abdominal tenderness epigastric
		789.67:abdominal tenderness general
		789.69:abdominal tenderness oth site
Alcoholism		303:alcohol dependence syndrome
		305.00:alcohol abuse-unspec
		305.01:alcohol abuse-continuous
		305.02:alcohol abuse-episodic
Allergic reaction		995.3:allergy, unspecified
		allergic reaction
		allergic rxn
Ankle fracture		824:fracture of ankle
Anticoagulated		790.92:abnormal coagulation profile
		e934.2:adv eff anticoagulants
		v58.61:long term use anticoagulant
		low molecular weight heparins
		anticoagulants – coumarin
		thrombin inhibitor – selective direct & reversible
		direct factor xa inhibitors
		vitamins – k, phytonadione and derivatives
		factor ix preparations
		factor ix complex (prothrombin complex concentrate) preparations
		ffp
asthma-copd		491:chronic bronchitis
		492:emphysema
		493:asthma
Back pain		724:other and unspecified disorders of back
Bicycle accident		e006.4:activities involving bike riding
		e800.3:rr coll nos-ped cyclist
		e801.3:rr coll w oth obj-cycl
		e802.3:rr acc w derail-ped cycl
		e803.3:rr acc w explos-ped cycl
		e804.3:fall from train-ped cycl

	🟩	e805.3:hit by train-ped cyclist
	🟩	e806.3:rr acc nec-ped cyclist
	🟩	e807.3:rr acc nos-ped cyclist
	🟩	e810.6:mv-train coll-ped cycl
	🟩	e811.6:reentrant coll-ped cycl
	🟩	e812.6:mv coll nos-ped cycl
	🟩	e813.6:mv-oth veh coll-ped cycl
	🟩	e814.6:mv coll w ped-ped cycl
	🟩	e815.6:mv coll w obj-ped cycl
	🟩	e816.6:loss control mv-ped cycl
	🟩	e817.6:mv brd/alight-ped cycl
	🟩	e818.6:mv traff acc-ped cyc
	🟩	e819.6:traffic acc nos-ped cycl
	🟩	e820.6:snow veh acc-ped cycl
	🟩	e821.6:oth off-road mv-ped cycl
	🟩	e822.6:oth coll mov obj-ped cyc
	🟩	e823.6:oth coll stn obj-ped cyc
	🟩	e824.6:n-traf brd/alit-ped cycl
	🟩	e825.6:mv n-traff nec-ped cycl
	🟩	e826:pedal cycle accident
		bicycle
		bike
Cancer	🟩	02:neoplasms
		adenocarcinoma
		aml
		breast cancer
		cancer
		carcinoma
		chemo
		chemotherapy
		cll
		hem onc
		hepatocellular
		hodgkin's
		hodgkins
		leukemia
		lumpectomy
		lung cancer
		lymphoma
		mastectomy

		melanoma
		metastasis
		metastatic
		mets
		multiple myeloma
		myeloma
		onc
		oncologist
		oncology
		radiation therapy
		remission
		stage iv
		xrt
Cardiac etiology	🟩	410:acute myocardial infarction
	🟩	411:other acute and subacute form of ischemic heart disease
	🟩	413:angina pectoris
	🟩	428:heart failure
	🟩	785.51:cardiogenic shock
	🟪	antianginal – coronary vasodilators (nitrates)
	🟪	diuretic – loop
	🟪	antianginal – coronary vasodilators (nitrates) combinations
Cellulitis	🟩	680-686:infections of skin and subcutaneous tissue
Chest pain	🟩	410:acute myocardial infarction
	🟩	411.1:intermed coronary synd
	🟩	413:angina pectoris
	🟩	786.50:chest pain nos
	🟩	786.59:chest pain nec
Congestive heart failure	🟩	428:heart failure
	🟪	diuretic – loop
Cholecystitis	🟩	574:cholelithiasis
	🟩	575.0:acute cholecystitis
Cerebrovascular accident	🟩	434:occlusion of cerebral arteries
	🟩	435:transcient cerebral ischemia
	🟩	436:cva
	🟩	437.8:cerebrovasc disease nec
	🟩	437.9:cerebrovasc disease nos
	🟪	thrombolytic – tissue plasminogen activators

Condition		Code / Term
Diabetes		250:diabetes mellitus
		diabetic therapy
Deep vein thrombosis		453.40:acute venous embolism and thrombosis of unspecified deep vessels of lower extremity
		453.41:acute venous embolism and thrombosis of deep vessels of proximal lower extremity
		453.42:acute venous embolism and thrombosis of deep vessels of distal lower extremity
		453.82:acute venous embolism and thrombosis of deep veins of upper extremity
		453.83:acute venous embolism and thrombosis of upper extremity, unspecified
Employee exposure		e920.5:hypodermic needle
		employee exposure
		needlestick
Epistaxis		784.7:epistaxis
Gastroenteritis		008:intestinal infections due to other organisms
		558:other noninfective gastroenteritis and colitis
		787.91:diarrhea
Gastrointestinal bleed		569.3:rectal & anal hemorrhage
		578:gastrointestinal hemorrhage
Headache		339:other headache syndromes
		346:migraine
		784.0:headache
Hematuria		599.70:hematuria, unspecified
		599.71:gross hematuria
		599.7:hematuria
Hiv+		042:hiv disease
		v08:asymptomatic hiv infection
		protease inhibitors (peptidic) antiretroviral
		protease inhibitors (non-peptidic) antiretroviral
		antiretroviral – ccr5 co-receptor antagonist
		antiretrovirals
		cd 4
		haart
		hiv
		hiv+
		430:subarachnoid hemorrhage

Intracerebral hemorrhage		431:intracerebral hemorrhage
		432:other and unspecified intracranial hemorrhage
		852:subarachnoid, subdural, and extradural hemorrhage, following injury
		853:other and unspecified intracranial hemorrhage following injury
Immunosuppressed		288.00:neutropenia, unspecified
		immunosuppressive agents
		glucocorticoids
		antineoplastic – antimetabolite – folic acid analogs
		anti-inflammatory – interleukin-1 beta blockers
		immunocompromised
		immunosuppressed
		+ All anchors from Cancer phenotype
Infection		01. infectious and parasitic diseases
		038:septicemia
		460-466:acute respiratory infections
		480-488:pneumonia and influenza
		540-543:appendicitis
		562.11:diverticulitis colon-no hem
		575.0:acute cholecystitis
		576.1:cholangitis
		590:infections of kidney
		595.0:acute cystitis
		599.0:urin tract infection nos
		680-686:infections of skin and subcutaneous tissue
		790.7:bacteremia nos
		995.91:sepsis
		995.92:severe sepsis
		cephalosporin antibiotics
		macrolide antibiotics and combinations
		glycopeptide antibiotics
		fluoroquinolone antibiotics
Kidney stone		592:calculus of kidney and ureter
		788.0:renal colic
Laceration		lac
		laceration
Liver (history)		571:chronic liver disease and cirrhosis
		572.2:hepatic encephalopathy

		cirrhosis
		esld
		hcv
		hep c
Motor Vehicle Accident	🟩	e810-e819:motor vehicle traffic accidents
From nursing home		nsg . home
		nsg facility
		nsg home
		nursing facility
		nursing home
Pancreatitis	🟩	577.0:acute pancreatitis
Pneumonia	🟩	480:viral pneumonia
	🟩	481:pneumococcal pneumonia
	🟩	482:other bacterial pneumonia
	🟩	483:pneumonia organism nec
	🟩	484:pneumonia in infectious diseases classified elsewhere
	🟩	485:broncopneumonia org nos
	🟩	486:pneumonia, organism nos
Psych	🟩	295:schizophrenic psychoses
	🟩	296:affective psychoses
	🟩	297:paranoid states
	🟩	298:other nonorganic psychoses
	🟩	311:depressive disorder nec
	🟩	v62.84:suicidal ideation
	🟩	v62.85:homicidal ideation
Obstruction	🟩	560.9:intestinal obstruct nos
Septic shock	🟩	785.52:septic shock
	🟪	cardiac sympathomimetics
Severe sepsis	🟩	785.52:septic shock
	🟩	995.92:severe sepsis
	🟪	cardiac sympathomimetics
Sexual assault	🟩	v71.5:observ following rape
Suicidal ideation	🟩	v62.84:suicidal ideation
		si
		suicidal ideation
Syncope	🟩	780.2:syncope and collapse
		syncopal episode
Uti	🟩	590:infections of kidney
	🟩	599.0:urin tract infection nos

Legend:

▦	Medication dispensing record	▥	Medication history
▨	ICD9 codes	☐	Medical Text

Electronic Medical Record Phenotyping using the Anchor & Learn Framework

Yoni Halpern, Steven Horng, Youngduck Choi, David Sontag

Phenotype Feature Weights

Phenotype	Data source	Observed Feature	Weight
Abdominal pain		pain	0.88
		abd	0.8
		abdo	0.72
		abdominal	0.64
		epigastric	0.59
		flank	0.58
		rlq	0.54
		abd	0.53
		rlq pain	0.53
		abd pain	0.51
		MetRONIDAZOLE (Flagyl)	0.5
		pain	0.48
		ct	0.47
		llq pain	0.46
		abdominal pain	0.46
		ruq	0.45
		abdominal	0.44
		llq	0.44
		sbo	0.43
		ercp	0.43
Alcoholism		etoh	2.71
		ETHANOL (189.5-273.5)	1.69
		etoh	1.44
		drinking	1.34
		ETHANOL (273.5-352.5)	1.23
		alcohol	1.19

TABLE 19. Electronic Medical Record Phenotyping using the Anchor & Learn Framework |Phenotype Feature Weights.

	🟩	sober	1.19
	🟨	drunk	1.11
	🟩	intoxicated	1.08
	🟦	ETHANOL (82.5-133.5)	1.06
	🟩	drinking	1.05
	🟦	ETHANOL (>352.5)	1.03
	🟨	detox	0.97
	🟨	drink	0.94
	🟩	drink	0.91
	🟨	intoxicated	0.89
	🟨	apparently	0.81
	🟨	drank	0.8
	🟨	alcohol	0.78
	🟦	ETHANOL(<82.5)	0.75
Allergic Reaction	🟪	DiphenhydrAMINE	1.43
	🟨	benadryl	1.13
	🟪	MethylPREDNISolone Sodium Succ	1.09
	🟪	DiphenhydrAMINE	1.05
	🟪	Famotidine	0.89
	🟩	benadryl	0.88
	🟨	neg:hives	0.86
	🟨	throat	0.79
	🟪	PredniSONE	0.73
	🟨	itching	0.72
	🟨	neg:sob	0.71
	🟨	swelling	0.7
	🟨	neg:rash	0.66
	🟪	Famotidine (PO)	0.63
	🟨	iv	0.63
	🟨	allergy	0.58
	🟨	feeling	0.52
	🟨	ate	0.52
	🟨	hives	0.51
	🟨	rash	0.51
Ankle Fracture	🟨	ankle	2.7
	🟩	ankle	1.27
	🟩	ortho	0.95
	🟨	fx	0.86
	🟨	fib	0.86

		Fentanyl Citrate	0.82
		fibula	0.68
		tib	0.64
		fall	0.62
		ankle pain	0.57
		fracture	0.53
		ankle injury	0.52
		fib	0.5
		swollen	0.47
		distal	0.47
		left ankle	0.46
		fx	0.45
		twisted	0.44
		HYDROmorphone (Dilaudid)	0.43
		splint	0.42
Anticoagulated		PT (>25.05)	1.95
		PT (21.05-25.05)	1.48
		coumadin	1.42
		INR(PT) (2.45-7.85)	1.4
		coumadin	1.34
		INR(PT) (2.05-2.45)	1.27
		INR(PT) (1.75-2.05)	1.24
		lovenox	1.19
		PTT (33.65-52.05)	1
		PT (19.15-21.05)	0.92
		amiodarone	0.91
		lovenox	0.87
		afib	0.84
		inr	0.82
		INR(PT) (1.65-1.75)	0.8
		warfarin	0.78
		PT (18.15-19.15)	0.75
		Lanoxin	0.72
		INR(PT) (1.45-1.65)	0.7
		PT (16.75-18.15)	0.65
Asthma-Copd		Albuterol 0.083% Neb Soln	1.89
		PredniSONE	1.79
		asthma	1.61
		MethylPREDNISolone Sodium Succ	1.33

		asthma	1.22
		copd	1.2
		albuterol sulfate	1.03
		Albuterol	0.82
		Ipratropium Bromide Neb	0.66
		copd	0.64
		Fluticasone-Salmeterol	0.61
		Albuterol Inhaler	0.59
		Advair Diskus	0.58
		sob	0.55
		wheezing	0.55
		wheezing	0.54
		albuterol sulfate	0.54
		nebs	0.52
		improved	0.5
		albuterol	0.5
Back pain		lbp	2.22
		back	1.49
		back pain	1.28
		lbp	1.15
		back	1.08
		sciatica	1.05
		sciatica	0.91
		lumbar	0.89
		spine	0.84
		coccyx	0.75
		low back	0.7
		mvc	0.68
		scapular	0.66
		thoracic	0.65
		flank	0.65
		back pain	0.64
		spasm	0.62
		lower back	0.54
		neg:trauma	0.54
		strain	0.52
Bicycle accident		bicyclist	1.86
		neg:helmet	1.41
		helmet	1.3
		abrasions	0.94

		neg:loc	0.88
		p fall	0.87
		chin	0.82
		bars	0.77
		handle	0.75
		car	0.72
		elbow	0.7
		20-30	0.64
		wrist	0.56
		neg:head	0.55
		loc	0.55
		abrasion	0.55
		fracture	0.53
		fractures	0.52
		head	0.49
		rib	0.49
Cancer (history)		omed	1.76
		ca	1.66
		Prochlorperazine Maleate	1.22
		dexamethasone	1.03
		ondansetron HCl	0.89
		acyclovir	0.88
		bmt	0.88
		ca	0.85
		radiation	0.67
		70-80	0.65
		breast	0.64
		60-70	0.59
		p resection	0.57
		prochlorperazine maleate	0.57
		Lorazepam	0.54
		anastrozole	0.53
		80-90	0.53
		fibroids	0.53
		resection	0.5
		Dexamethasone Sod Phosphate	0.5
Cardiac Etiology		nstemi	2.22
		stemi	1.55
		.	0.84
		echo	0.73

		zoll	0.68
		cath	0.6
		cTropnT (0.045-0.365)	0.54
		ccu	0.52
		ekg	0.41
		st	0.38
		cmed	0.37
		neg:heparin	0.35
		Clopidogrel	0.35
		Atropine Sulfate	0.34
		Aspirin	0.34
		CK(CPK) (89.5-1111.5)	0.34
		trop	0.34
		bradycardia	0.33
		ste	0.33
		heparin	0.33
Cellulitis		cellulitis	2.16
		Vancomycin	1.93
		cellulitis	1.83
		Cephalexin	1.72
		abcess	1.71
		abscess	1.49
		CefazoLIN	1.48
		cyst	1.37
		unasyn	1.31
		Sulfameth/Trimeth DS	1.3
		infection	0.98
		redness	0.98
		axilla	0.92
		infected	0.88
		vanco	0.87
		i	0.86
		erythema	0.82
		finger	0.8
		erythema	0.8
		LACTATE (1.03-1.55)	0.76
Chest pain		cp	1.78
		chest	1.33
		cp	1.14
		chest	0.94

	🟨	chest pain	0.83
	🟪	Aspirin	0.75
	🟪	Aspirin	0.71
	🟨	rib	0.66
	🟦	CK(CPK) (100.5-712.5)	0.57
	🟪	Nitroglycerin SL	0.55
	🟦	CK(CPK) (62.5-100.5)	0.54
	🟩	sets	0.53
	🟩	chest pain	0.53
	🟨	sscp	0.53
	🟨	rib pain	0.52
	🟩	stress	0.5
	🟦	CK(CPK) (41.5-56.5)	0.45
	🟩	cxr	0.42
	🟪	Aspirin (Buffered)	0.42
	🟪	Clopidogrel	0.41
Congestive Heart Failure	🟩	lasix	2.28
	🟦	proBNP (7827.5-21173.0)	1.82
	🟦	proBNP (1703.0-7510.0)	1.76
	🟦	proBNP (940.0-1507.0)	1.48
	🟩	chf	1.26
	🟦	proBNP (>21490.0)	1.24
	🟪	Nitroglycerin	1.09
	🟨	chf	1.04
	🟪	Nitroglycerin Ointment 2%	0.96
	🟪	Mannitol 20%	0.96
	🟩	overload	0.87
	🟪 (magenta)	Furosemide	0.84
	🟪	Lisinopril	0.82
	🟪 (magenta)	Lasix	0.77
	🟨	sob	0.76
	🟦	proBNP (1507.0-1663.5)	0.74
	🟦	proBNP (270.5-940.0)	0.73
	🟪	Docusate Sodium	0.69
	🟪 (magenta)	Lasix	0.68
	🟩	bnp	0.61
Cholecystitis	🟩	gallstones	1.09
	🟩	cholelithiasis	1.07
	🟨	gallstones	1.04
	🟩	cholecystitis	0.94

		ruq	0.89
		LIPASE(<430.5)	0.84
		us	0.79
		surgery	0.77
		unasyn	0.75
		us	0.73
		ercp	0.67
		surg	0.65
		ruq	0.65
		MetRONIDAZOLE (Flagyl)	0.64
		pain	0.63
		ALK_PHOS (152.5-164.5)	0.63
		cholecystitis	0.59
		stone	0.59
		neg:cholecystitis	0.59
		stone	0.58
Cerebral Vascular Accident		stroke	1.11
		cva	1.1
		neuro	0.97
		infarct	0.93
		stroke	0.81
		tia	0.79
		weakness	0.73
		Heparin Sodium	0.69
		tia	0.68
		resolved	0.56
		infarct	0.55
		mca	0.55
		mri	0.51
		tpa	0.47
		droop	0.47
		neurology	0.46
		admit to neuro	0.46
		PT (13.35-27.15)	0.45
		asa	0.45
		80-90	0.45
Diabetes (history)		dm	2.97
		Ascensia Contour	2.92
		dm2	2.23
		GLUCOSE (>266.5)	2.1

		MetFORMIN (Glucophage)	1.98
		iddm	1.87
		GLUCOSE (198.5-266.5)	1.8
		dmii	1.72
		diabetes	1.56
		FreeStyle Lancets	1.47
		diabetic	1.42
		Ascensia Contour	1.25
		diabetic	1.22
		hypoglycemia	1.22
		iddm	1.19
		bs	1.16
		Insulin HumaLOG	1.16
		GLUCOSE (175.5-198.5)	1.13
		Tricor	1.1
		dm1	1.1
Deep Vein Thrombosis		dvt	1.94
		dvt	1.43
		Heparin Sodium	1.21
		Warfarin	1.14
		Enoxaparin Sodium	0.96
		leg	0.79
		Heparin Sodium	0.75
		lovenox	0.73
		swelling	0.71
		calf	0.64
		left leg	0.56
		PT (11.65-15.45)	0.55
		INR(PT)(<1.25)	0.53
		rle	0.5
		filter	0.46
		lle	0.46
		us	0.46
		anticoagulation	0.45
		clot	0.45
		heparin	0.44
Employee Exposure		needle	1.9
		TriagePain(<0.05)	1.47
		LaMIVudine-Zidovudine (Combivir)	1.41
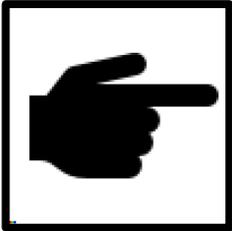
	🟨	or	1.36
	🟨	stuck	1.13
	🟨	exposure	1.06
	🟨	neg:bleeding	1
	🟨	washed	0.98
	🟨	went	0.96
	🟢	TriageTemp (98.98-99.21)	0.95
	🟨	cath	0.94
	🟨	epi	0.93
	🟨	glove	0.91
	🟨	dirty	0.81
	🟨	sq	0.8
	🟨	thumb	0.77
	🟩	patient	0.77
	🟩	needle	0.73
	🟢	TriageHR (61.5-66.5)	0.72
	🟩	id	0.72
Epistaxis	🟨	epistaxis	6.15
	🟨	nose	2.43
	🟩	epistaxis	2.37
	🟨	nosebleed	1.79
	🟪	Oxymetazoline 0.05%	1.55
	🟨	bleed	1.32
	🟨	nares	1.12
	🟨	arrest	1.04
	🟨	cardiac	0.98
	🟩	ent	0.9
	🟨	nare	0.89
	🟨	appears	0.84
	🟢	TriageHR (105.5-107.5)	0.82
	🟨	neg:controlled	0.8
	🟨	swollen	0.79
	🟨	neg:thinners	0.77
	🟢	TriagePain(<0.05)	0.73
	🟩	face	0.73
	🟨	neg:blood	0.71
	🟨	spontaneous	0.7
Gastroenteritis	🟨	diarrhea	2.54
	🟩	diarrhea	1.91
	🟨	d	1.53

		diarhea	1.11
		gastroenteritis	1.1
		diff	1.07
		stool	1.04
		diahrrea	1.02
		colitis	0.98
		nvd	0.79
		stool	0.73
		diff	0.73
		immodium	0.68
		Ondansetron	0.68
		MetRONIDAZOLE (Flagyl)	0.68
		abdo	0.62
		sick	0.6
		n	0.57
		nvd	0.56
		loose	0.56
Gastrointestinal Bleed		Pantoprazole Sodium	2.35
		brbpr	2.34
		gi	1.64
		rectal	1.44
		brbpr	1.4
		rectal bleeding	1.09
		blood	1.04
		stool	0.99
		gib	0.94
		hct	0.91
		bloody	0.9
		hematemesis	0.8
		guaiac	0.79
		gi bleed	0.79
		Lidocaine Jelly 2% (Urojet)	0.74
		blood in stool	0.72
		gib	0.68
		stools	0.62
		blood	0.62
		brbpr x	0.62
Headache		headache	2.25
		ha	2.2
		h	1.87

		migraine	1.72
		ha	1.58
		headache	1.5
		Prochlorperazine	1.34
		head	1.09
		WBC(<7.5)	1.05
		migraine	1.01
		photophobia	0.96
		migraines	0.95
		Acetaminophen-Caff-Butalbital	0.93
		headaches	0.93
		headaches	0.78
		migraines	0.76
		neg:vision	0.71
		Prochlorperazine Maleate	0.7
		head	0.61
		esr	0.58
Hematuria		hematuria	3.34
		hematuria	1.8
		urine	0.83
		clots	0.83
		urology	0.79
		blood	0.74
		foley	0.74
		EPI(<0.5)	0.62
		bleeding	0.48
		neg:dysuria	0.47
		foley	0.46
		abd	0.44
		Lidocaine Jelly 2% (Urojet)	0.44
		bladder	0.41
		Male	0.41
		PH (8.25-8.75)	0.4
		noticed	0.39
		penile	0.39
		SP_GRAV (1.0136-1.015)	0.39
		blood	0.38
HIV+ (history)		Truvada	4.84
		ATRIPLA	3.61
		Epzicom	2.68

		Tenofovir Disoproxil Fumarate	2.18
		Combivir	2.01
		Lamivudine	1.8
		cd4	1.55
		Raltegravir	1.5
		id	1.18
		abacavir	1.12
		Kaletra	0.97
		40-50	0.96
		Sustiva	0.92
		Etravirine	0.89
		Epivir	0.88
		exposure	0.83
		Viramune	0.76
		vl	0.76
		Dapsone	0.73
		azithromycin	0.73
Intracerebral hemorrhage		sdh	2.11
		sdh	1.6
		ich	1.51
		sah	1.43
		bleed	1.42
		neurosurg	1.23
		head bleed	1.17
		sah	1.06
		bleed	0.97
		nsurg	0.94
		subdural	0.93
		neurosurgery	0.91
		ich	0.84
		hemorrhage	0.77
		Labetalol	0.74
		ct	0.71
		Phytonadione	0.69
		headache	0.66
		subdural	0.64
		shift	0.61
Immunosuppressed		Truvada	2.84
		ATRIPLA	1.91
		Bactrim	1.88

		omed	1.61
		ca	1.33
		hydroxychloroquine	1.21
		cd4	1.2
		prednisone	1.15
		Epzicom	1.1
		Tenofovir Disoproxil Fumarate	1.01
		Raltegravir	0.99
		transplant	0.97
		Prochlorperazine Maleate	0.94
		prednisone	0.87
		ondansetron HCl	0.82
		Combivir	0.77
		abacavir	0.7
		bmt	0.7
		Bactrim DS	0.68
		acyclovir	0.67
Infection		cipro	1.66
		vanc	1.44
		ceftriaxone	1.33
		MetRONIDAZOLE (Flagyl)	1.28
		uti	1.25
		Sulfameth/Trimeth DS	1.21
		azithro	1.07
		ancef	1.06
		pna	1.05
		levaquin	1.04
		vanco	1.03
		abx	1.03
		cellulitis	1.03
		keflex	0.97
		cellulitis	0.93
		st	0.93
		vancomycin	0.88
		azithromycin	0.83
		MetRONIDAZOLE (Flagyl)	0.81
		cough	0.78
Kidney stone		Ketorolac	1.58
		stone	1.38

	🟨	flank	1.24
	🟨	stone	1.18
	🟨	pain	0.93
	🟩	stones	0.92
	🟩	urology	0.92
	🟦	SP_GRAV (1.0155-1.0175)	0.77
	🟦	PH (6.25-6.75)	0.77
	🟨	stones	0.73
	🟦	EPI(<0.5)	0.69
	🟨	flank pain	0.67
	🟦	PH(<5.25)	0.63
	🟦	PH (6.75-7.25)	0.6
	🟨	kidney	0.58
	🟩	nephrolithiasis	0.57
	🟦	SP_GRAV (1.0185-1.0265)	0.55
	🟦	CREAT (1.35-1.65)	0.53
	🟦	PROTEIN(<27.5)	0.52
	🟩	mm	0.51
Laceration	🟪	Tetanus-Diphtheria Tox (DECAVAC)	1.6
	🟪	Lidocaine 1%/Epi 1::100,000	1.28
	🟨	glass	1.21
	🟩	suture	1.19
	🟨	controlled	1.18
	🟨	tetanus	1.16
	🟨	knife	1.03
	🟨	dsd	0.97
	🟨	sutures	0.86
	🟨	plastics	0.85
	🟨	bleeding controlled	0.84
	🟨	sutured	0.82
	🟩	sutures	0.81
	🟩	tetanus	0.81
	🟨	cut	0.77
	🟩	cut	0.73
	🟨	cutting	0.7
	🟨	applied	0.69
	🟨	box	0.68
	🟨	trying	0.68
Liver disease	🟪	Xifaxan	0.96

[magenta]	spironolactone	0.89
[green]	liver	0.86
[magenta]	Ribavirin	0.82
[magenta]	Spironolactone	0.8
[magenta]	ribavirin	0.73
[green]	ascites	0.73
[magenta]	Truvada	0.73
[magenta]	lactulose	0.69
[purple]	OxycoDONE (Immediate Release)	0.69
[cyan]	AST(SGOT) (80.5-213.5)	0.68
[cyan]	ALT(SGPT) (89.5-114.5)	0.67
[green]	o hiv	0.67
[red]	50-60	0.67
[green]	heroin	0.63
[cyan]	PLT_COUNT (53.5-62.5)	0.61
[yellow]	needle	0.59
[cyan]	PLT_COUNT (62.5-89.5)	0.59
[red]	40-50	0.57
[magenta]	nadolol	0.57
Motor vehicle accident [yellow]	mvc	3.68
[yellow]	motorcycle	1.8
[yellow]	mva	1.63
[yellow]	mcc	1.6
[green]	mvc	1.41
[yellow]	car	1.28
[yellow]	ped struck	1.28
[yellow]	ped	1.25
[yellow]	accident	1.22
[green]	mcc	0.91
[yellow]	passenger	0.82
[yellow]	struck	0.8
[yellow]	restrained	0.79
[yellow]	fx	0.77
[yellow]	scooter	0.74
[yellow]	bus	0.68
[yellow]	driver	0.65
[yellow]	mph	0.65
[yellow]	rearended	0.63
[yellow]	pedestrian	0.62

Nursing home		nh	0.53
		90+	0.51
		80-90	0.48
		Mapap (acetaminophen)	0.44
		staff	0.44
		bisacodyl	0.43
		dementia	0.4
		baseline	0.39
		TriagePain (12.75-13.5)	0.34
		70-80	0.33
		Vancomycin	0.32
		arrives via ems	0.3
		nh	0.27
		Colace	0.27
		Phillips Milk of Magnesia	0.26
		tube	0.26
		,	0.26
		dementia	0.25
		trazodone	0.24
		senna	0.24
Pancreatitis		pancreatitis	1.7
		pancreatitis	1.61
		LIPASE (>1859.5)	1.58
		LIPASE (618.0-1859.5)	1.2
		LIPASE (108.5-184.5)	1.19
		TOT_BILI (1.15-5.55)	0.87
		ALK_PHOS (115.5-266.5)	0.84
		TOT_BILI(<1.15)	0.83
		LIPASE (388.5-550.0)	0.82
		lipase	0.74
		LIPASE (192.5-388.5)	0.73
		ALT(SGPT) (22.5-74.5)	0.63
		epigastric	0.59
		promethazine	0.57
		LIPASE (184.5-192.5)	0.56
		AST(SGOT) (323.5-576.0)	0.52
		ALK_PHOS(<91.5)	0.52
		ALK_PHOS (91.5-115.5)	0.5
		abd	0.46
		pancreas	0.45

Pneumonia		Levofloxacin	1.68
		pna	1.67
		pneumonia	1.41
		Levofloxacin	1.19
		Azithromycin	1.07
		Levofloxacin	0.97
		CeftriaXONE	0.8
		pna	0.77
		infiltrate	0.69
		Vancomycin	0.65
		Azithromycin	0.65
		cough	0.64
		pneumonia	0.64
		rll	0.61
		lll	0.6
		opacity	0.57
		cxr	0.56
		cough	0.51
		LACTATE(<1.395)	0.45
		azithro	0.43
Pyschiatric Disorder		psych	2.02
		si	2.01
		depression	1.97
		si	1.07
		paranoid	0.97
		bipolar	0.87
		neg:si	0.87
		depression	0.84
		psych	0.83
		depressed	0.8
		schizophrenia	0.74
		plan	0.72
		confusion	0.71
		manic	0.69
		Xanax	0.62
		schizoaffective	0.61
		psychiatry	0.6
		LITHIUM (0.45-0.85)	0.59
		bournewood	0.59
		psychosis	0.59

Obstruction	🟩	sbo	1.77
	🟪	Lidocaine Jelly 2% (Urojet)	1.66
	🟨	sbo	1.48
	🟩	obstruction	0.92
	🟩	kub	0.84
	🟩	ngt	0.77
	🟩	surgery	0.71
	🟪	Ondansetron	0.68
	🟨	obstruction	0.67
	🟩	surg	0.64
	🟨	abd	0.62
	🟥	70-80	0.53
	🟩	ng	0.52
	🟨	bowel	0.5
	🟩	bowel	0.47
	🟩	partial	0.46
	🟨	v	0.43
	🟨	ngt	0.42
	🟩	lactate	0.42
	🟨	neg:output	0.41
Septic shock	🟩	levophed	0.93
	🟩	line	0.66
	🟪	Fentanyl Citrate	0.65
	🟩	hypotensive	0.51
	🟩	hypotension	0.51
	🟩	ij	0.45
	🟩	cvl	0.45
	🟪	Vancomycin	0.4
	🟪	Midazolam	0.39
	🟩	central	0.38
	🟩	placed	0.37
	🟩	icu	0.35
	🟦	LACTATE (1.85-2.55)	0.34
	🟦	LACTATE (4.95-7.15)	0.32
	🟩	pressors	0.32
	🟩	dopamine	0.32
	🟦	PH (6.71-7.075)	0.3
	🟦	PCO2 (34.5-50.5)	0.27
	🟪	Midazolam	0.27
	🟦	cTropnT (0.415-3.71)	0.27

Severe sepsis		levophed	1.19
		Midazolam	1.06
		Vancomycin	0.91
		cvl	0.87
		line	0.83
		hypotensive	0.77
		TriageSBP (60.5-79.5)	0.7
		K+ (2.65-3.15)	0.66
		placed	0.66
		icu	0.64
		Atropine Sulfate	0.61
		pressors	0.61
		hypotension	0.6
		LACTATE (1.85-3.05)	0.6
		LACTATE (5.75-12.65)	0.6
		Midazolam	0.57
		BASE_XS (-23.5--6.5)	0.54
		arrest	0.54
		Glucagon	0.5
		urology	0.48
Sexual assault		rci	3.3
		assault	1.79
		sane	1.65
		CeftriaXONE	1.52
		Azithromycin	1.5
		sexual	1.24
		MetRONIDAZOLE (Flagyl)	1.19
		Prochlorperazine Maleate	1.04
		LaMIVudine-Zidovudine (Combivir)	0.97
		presents	0.75
		rci	0.68
		sexually	0.6
		Zyrtec	0.58
		assault	0.58
		Ondansetron ODT	0.53
		Lidocaine 1%	0.52
		SP_GRAV(<1.0035)	0.51
		TriageDBP (82.5-83.5)	0.5
		ALT(SGPT) (15.5-21.5)	0.5

151

	🟩	TriageSBP (139.5-143.5)	0.46
Suicidal ideation		psych	2.36
		plan	1.4
		himself	1.07
		psychiatry	0.98
		neg:plan	0.98
		neg:plan	0.95
		suicidal	0.91
		plan	0.76
		depression	0.72
		self	0.71
		suicide	0.71
		kill	0.7
		sitter	0.67
		Nicotine Patch	0.66
		od	0.64
		cutting	0.63
		cutting	0.58
		neg:hi	0.55
		jump	0.55
		boyfriend	0.55
Syncope		syncope	2.89
		syncope	2.8
		syncopal	1.71
		passed	1.53
		syncopal	1.49
		loc	1
		presyncope	0.95
		fainted	0.94
		syncopized	0.9
		sycope	0.86
		lightheaded	0.78
		out	0.74
		dizziness	0.69
		unresponsive	0.68
		fainting	0.67
		hot	0.61
		faint	0.6
		vagal	0.6
		consciousness	0.59

		loc	0.58
Urinary tract infection		Ciprofloxacin	2.94
		Ciprofloxacin	2.34
		Ciprofloxacin IV	2.25
		uti	2.03
		Sulfameth/Trimeth DS	1.86
		uti	1.42
		CeftriaXONE	1.22
		dysuria	1.15
		pyelo	1.07
		PH (6.25-6.75)	1.02
		SP_GRAV (1.0109-1.0175)	1
		PROTEIN (52.5-87.5)	0.96
		positive	0.94
		PH(<5.25)	0.93
		PH (7.75-8.25)	0.92
		PH (6.75-7.25)	0.9
		PROTEIN (125.0-225.0)	0.88
		PH (5.75-6.25)	0.88
		PROTEIN (>400.0)	0.85
		SP_GRAV (1.0085-1.0109)	0.84

Legend:

Triage assessment   MD comments   Medication History   Medication Dispensing Record

Sex   Triage vitals   Lab results   Age

Anchor Finding Interface

V1.0

Yoni Halpern

November 17, 2014

Hi! This document accompanies the initial release of the anchor interface described in: "Using Anchors to Estimate Clinical State without Labeled Data" by Y. Halpern, Y.D. Choi, S. Horng, D. Sontag. To appear in the American Medical Informatics Association (AMIA) Annual Symposium, Nov. 2014.

Contact Information

Please direct questions to Yoni Halpern: halpern@cs.nyu.edu

Figure11. Anchor Finding Interface v1.0 | Anchor Elicitation Tool Working Screen Shot
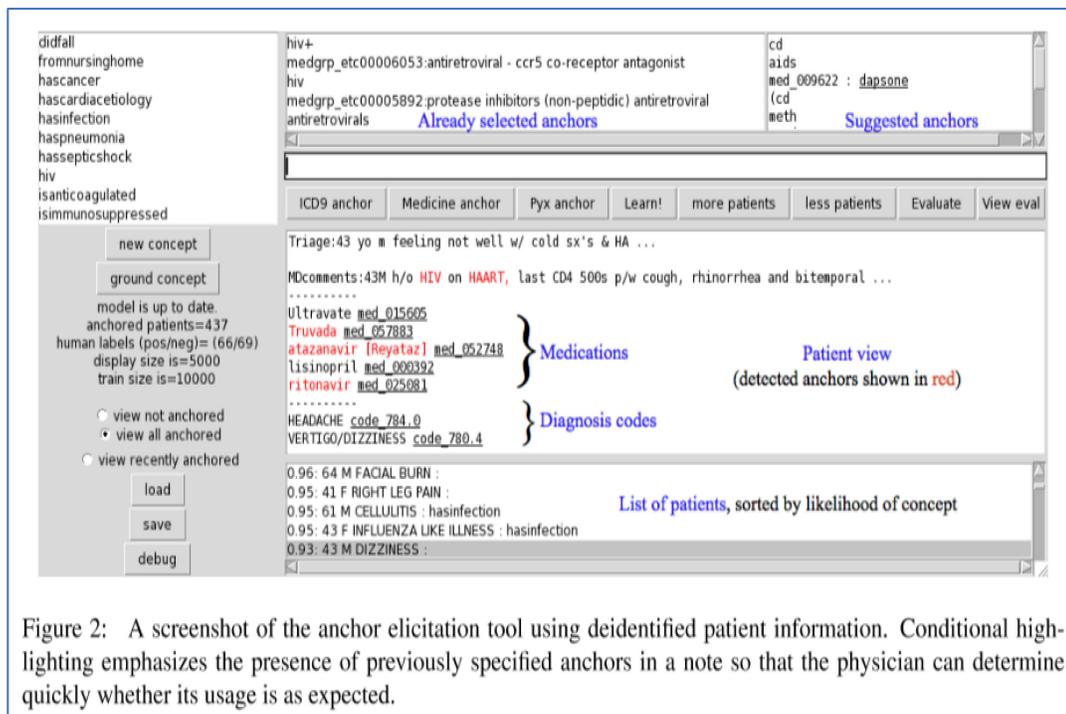
Figure 2:   A screenshot of the anchor elicitation tool using deidentified patient information. Conditional highlighting emphasizes the presence of previously specified anchors in a note so that the physician can determine quickly whether its usage is as expected.

QuickStart

1.1 Installation

Clone the GitHub repository for AnchorExplorer with the following command:

```
$ git clone https://github.com/yhalpern/anchorExplorer.git
$ cd anchorExplorer
```

We have tested the interface on a system with the following properties:

- Mac or unix OS
- python 2.6 or 2.7
- numpy 1.8.1 and scipy 0.13.3
- scikit-learn 0.15
- networkx 1.8.1
- Tkinter Revision 81008
- ttk 0.3.1

Required python packages are listed a file, requirements.txt and can be installed using pip:

```
$ pip install -r requirements.txt
```

If you have trouble with Tkinter and ttk, try installing the ActiveState community edition of Python http://www.activestate.com/activepython/.

1.2 Data

To start, you need to have patient records in xml format. However, if you want to test that you can get things up and running with dummy data, you can use the following command to generate 1000 properly formatted "random" patient records and store them in patients.xml. A second, more involved example can be found in Section 4.3.

```
$ python generate_patients.py 1000 > patients.xml
```

1.3 Settings

An example settings file is provided in the examples/ directory. To customize this file for your own data, see section 4.3.

1.4 Preprocess Data

Use the preprocess patients.py script to preprocess the data in patients.xml for easy lookups and storage. The interface will read from the files generated by this script. Using the example settings file:

```
$      python      preprocess_patients.py      1000      patients.xml
examples/settings.xml
```

This script does some simple negation detection, bigram detection, and stopword removal.

If you wish to use your own custom language processing pipeline, see section 4.2.

1.5 Createananchorsdirectory

```
$ mkdir  anchors
```

1.6 Runtheinterface

To run the interface using the example settings file:

```
$ python  gui.py examples/settings.xml
```

Figure12. Anchor Finding Interface v1.0 | Anchor Elicitation Tool Working Screen Shot.

Figure 12: (Left) Screenshot when the interface is first activated. Panes are annotated in red, and each one is explained in detail in Section 3. (Right) Viewing a patient in more detail. Random words are inserted for patient texthere.
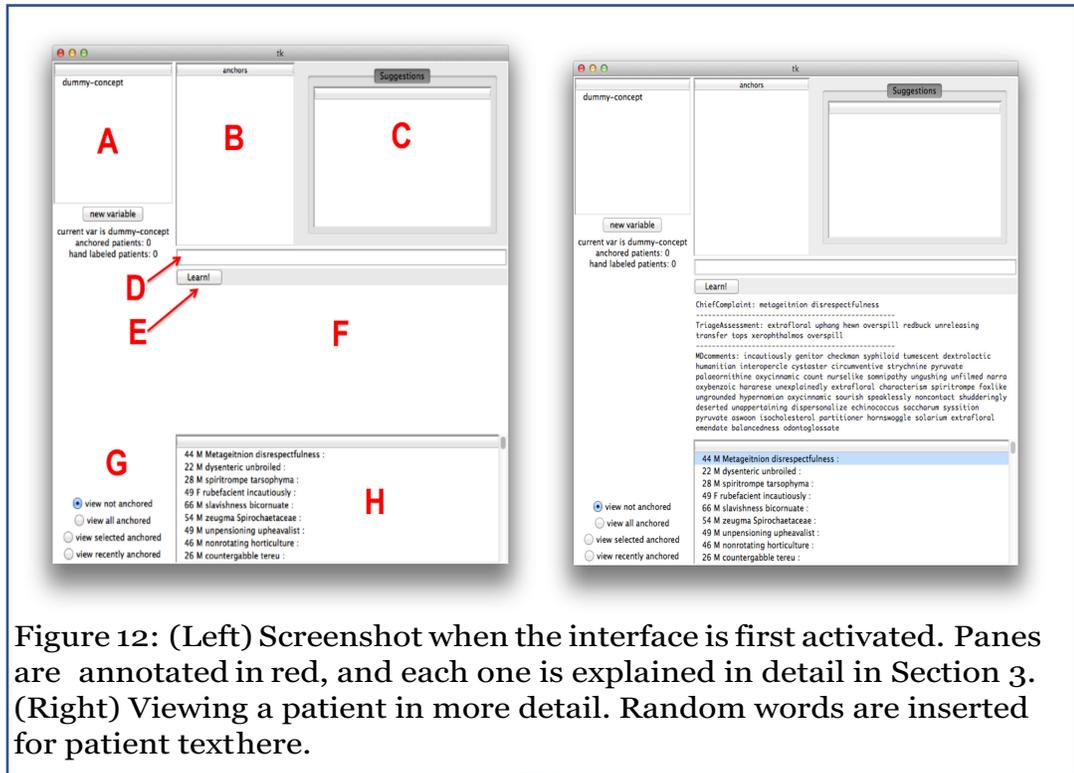
1.7 Exploring the data

You can explore the data on the patient visit level with the patient summary display[1] (H) and detailed display (F). Selecting a patient in the summary display shows the patient's record in more detail above it. In this simple example, we are only showing three text fields, but more data can be shown here.

1.8 Create a new cohort

Pushing the "new variable" button (below A) will pop up a dialog to create a new cohort variable. You will be prompted to provide a name.

1.9 Adding an anchor

Choose a word that appears in the patient data (e.g., for the randomly generated data, try "bioluminescence") and enter it in the text input box (D).

---

[1]This display sometimes gets initialized with a height of zero at the bottom of the window and may need to be dragged up in order to be visible.

1.10 Filtering the data

Choose to filter for patients that have the anchor or view only patients that do not have the anchor by using the radio buttons to change the filters (G).

1.11 Learning a model

Until now, you have just been exploring the patient data and applying filters, but no learning has taken place. Now try pushing the "Learn" button (E). After a few seconds, your patients should now be ranked in order of likelihood of belonging to the cohort, and suggestions for additional anchors will appear in the suggestions pane (C).

1.12 Saving state

Try closing the window and opening it up again. Your cohorts and anchors should be preserved.

2 Anatomy of the interface

2.1 Cohort Selection Menu (A)

Allows you to select cohorts, and create new ones. Some commands:

- New variable button to create a new cohort

- '-' delete cohort

- 'r' rename cohort

2.2 Anchor Display (B)

Displays the current anchors for the selected topic. Commands:

- Enter text into text box and push enter to add a new anchor

- '-' delete selected anchor

- Anchors can also be added using the anchor suggestion pane (Top Right)

2.3 Anchor Suggestion Pane (C)

Displays suggestions for anchors and allows for navigation and adding of structured anchors. Commands:

- "+" to add selected anchor suggestion

Important: The anchor suggestions are not necessarily all good - choosing anchors requires judgment of whether to accept or reject the automated suggestions. They are meant to trigger associations, not tell you what makes a good anchor or not!

2.4 Anchor Input Box (D)

Type anchors in here, push enter to apply them. You will need to push learn to relearn the model after adding an anchor.

2.5 Learn Button (E)

Learns a model using the currently selected anchors.

2.6 Detailed Patient Display (F)

Detailed display of the patient selected in Patient List (Bottom Right). Anchors are highlighted red. Structured data types can be clicked to navigate directly to the relevant part of the hierarchy (displayed in Anchor Suggestion Pane).

2.7 Patient Filters (G)

Possible options here:

- view not anchored: patients that do not have an anchor.

- view all anchored: all patients that have an anchor.

- view selected anchor: view patients that have the selected anchor (selected in pane B).

- view recent anchor: view patients added by the most recent anchor.

2.8 Patient List (H)

List of patients with summary information. These patients can be filtered using the radio buttons on the left (e.g., view all anchored, view not anchored, etc.). After the concept has been "learned" using the Learn Button, this list will be ranked in order of how highly the patient fits the currently selected concept.
Usage:

- arrow keys to navigate
- "+" mark patient as a positive example
- "-" mark patient as a negative example
- "0" remove patient marking

3.0 Customizations

This section is still incomplete; more information coming here shortly. Please contact if you have questions about how to use this tool on your data.

3.1 patientSets

This section describes which patients should be selected for visualization. If you have many patients, the interface may run slowly, here you can show that you only want to use the first 10,000 patients for an initial exploration.

```
1. <patientSets>
2.     <set name='train' start='0' end='10000'/>
3.     <set name='validate' start='0' end='15000'/>
4. </patientSets>
```

The only important patientSet in this version of the code are the "train" and "validate" set. The start and end fields indicate indices of patients as they are listed in the file visitIDs, which is generated by the preprocessing script. Only patients from the validate set are shown. Patients in the train set are not shown.

3.2 Language Parsing

Coming soon... more information on customizing the language parser.

3.3 Data Fields

Use the settings.xml file to customize the interface for your particular dataset. An example settings.xml file is provided along with the interface

159

code in the examples directory. Most of it can be left as is. The most important parts to customize are the dataTypes and displaySettings sections.

3.4 Data Types

We divide data into different *types* (e.g., text, age, sex, medications, diagnoses, procedures, etc.). Each type is denoted by a datum tag and contains a number of field tags that describe where this data appears in a patient xml representation.

For example, here is a sample patient representation in xml format generated by the random patient generator:

```
//patient.xml
<visit>
   <index>qVwLYjLKqlkZhvkf</index>
   <MDcomments> SOME TEXT</MDcomments>
   <Age>  44 </Age>
   <Sex>M</Sex>
   <ChiefComplaint> SOME  TEXT</ChiefComplaint>
   <TriageAssessment>SOME TEXT</TriageAssessment>
</visit>
```

Important: Every patient instance must be enclosed in a visit tag and must have a unique index tag. All other fields are customizable. Here is a section of settings.xml that could be used for patient records with this   schema.

```
//settings.xml
1. <datum type='text' heirarchy=''  prefix=''>
2.       <field name='MDcomments' path='.'/>
3.       <field  name='ChiefComplaint'  path='.'/>
4.       <field  name='TriageAssessment'  path='.'/>
5. </datum>
```

Line 1 of settings.xml says that there is a type of data called **text**. That it is not hierarchical and that it should not be represented with any special prefix.

Lines 2-4 show where in the patient records it can be found (i.e., the MDcomments, ChiefComplaint, and TriageAssessment sections). The final representation of the **text** data will be a concatenation of these three fields.

3.5 Display Settings

For each type of data described in Section 4.3.1, you may want to customize

where it is displayed (or whether it is displayed at all). Here you can specify what appears in the patientSummary section and the detailedDisplay.

The following snippet says that Age, Sex and ChiefComplaint should be displayed as the patient summary:

```
<patientSummary>
    <displayFields>
        <field  name='Age'/>
        <field name='Sex'/>
        <field name='ChiefComplaint'/>
    </displayFields>
</patientSummary>
```

And the following snippet says that ChiefComplaint, TriageAssessment, MDcomments should be displayed in the detailed patient description:

```
<detailedDisplay>
    <displayFields>

        <field  name='ChiefComplaint' path='.'/>
        <field  name='TriageAssessment'  path='.'/>
        <field name='MDcomments' path='.'/>
    </displayFields>
</detailedDisplay>
```

The resulting display is shown in Figure 2.

Figure13. Anchor Finding Interface v1.0 | Anchor Elicitation Tool Customize Patient Display.
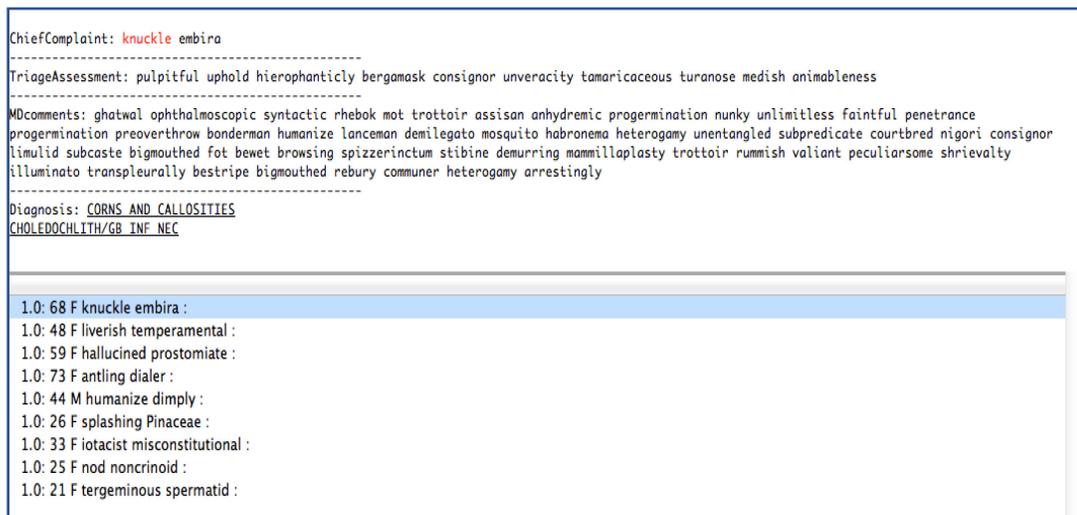
Figure 13: Customizing the patient display as described in section 4.3.2. The bottom list view shows Age, Sex, and Chief Complaint for every patient. The top detailed view of a single patient shows fields: ChiefComplaint, TriageAssessment,MDcomments. We display random words instead of real patient notes.

4.4 Structured Data Types

Some data elements can be described as belonging to a hierarchy of items, (e.g., ICD9 codes, NDC codes, etc.). We support exploring data hierarchies and adding anchors from hierarchically typed data. In the following section, we demo this capability in order to view diagnosis codes in MIMIC data alongside text.

4.5 Customization demo

This demo will walk you through how to include a new data field; in this case, ICD9 codes to be viewed in the user interface. It assumes that you have access to MIMIC-II data.

First,   download   the   flat   note   files   from http://physionet.org/mimic2/flat_files/.

4.5.1 Preparing data

The commands from this demo can be found in a script mimicdemo.sh, which takes a single argument, which is the path to the mimic-II flat files.
As before, we begin by formatting the patient data in a patients.xml file.

```
$ python get_mimic_data.py path/to/mimic_files/00
examples/mimic_fields.txt
> patients.xml
```

Looking at the patients.xml file, you will see that ICD9 codes are recorded in the dataset.

We would like to create a *data dictionary* to map these codes to plaintext and a *data structure* to store these codes in a hierarchical manner.

First, we need two tab-separated files that will store this data: X.names and X.edges. In general, we assume that the user supplies   these   files   for   any structured   data   type   of   interest. For this example, we provide a script examples/ICD9/load ICD9 structure.py, which downloads the ICD9 hierarchy from a public git repository and creates these formatted files. **This hierarchy is incomplete and does not contain entries for every ICD9 code in MIMIC-II, but it serves as a good illustration**.

162

```
$ cd examples/ICD9
$ python load_ICD9_structure.py
```

We now want to build these into the structured format that the interface reads from.

```
$ cd ../..
$ python build_structured_rep.py code examples/ICD9/code
$ ls Structures
codeDict.pk     codeStruct.pk
```

Now we need to update settings.xml to recognize this new structured data type. An example settings file is found in examples/ICD9/settings.xml. Below is an excerpt of the relevant lines:

```
#lines 28-30
 <datum type='code' heirarchy='Structures/codeStruct.pk' prefix='code_'
  dictionary='Structures/codeDict.pk'  realtime='true'>
    <field  name='Diagnosis' path='D_code'  display='D_name'/>
 </datum>


 #lines 41-45
 <displayFields>
    <field  name='ChiefComplaint' path='.'/>
    <field  name='TriageAssessment'  path='.'/>
    <field name='MDcomments' path='.'/>
    <field  name='Diagnosis' path='.'/>
  </displayFields>
```

We can now run the preprocessing script to import the data with the structured ICD9 codes included.

```
$ python preprocess_patients.py 1000 patients.xml
examples/ICD9/settings.xml
```

And run the interface:

```
$ python gui.py  examples/ICD9/settings.xml
```

4.5.2 Using the interface

Try to create a new cohort and add an anchor: As an illustrative example, we will build the cohort of recent births.

Create a new cohort:
Push the **new variable** button (Pane A in Figure 13) and enter "recent birth."

Adding a bad anchor:

We'll start with a straw-man bad anchor for recent birth. Say we hypothesize that the word "mother" is a good anchor for recent birth (this would mean that if mother is mentioned in the note, it is almost certainly a recent birth. This is almost certainly not true, but let's run with it and see what happens.)

Type "mother" in the anchor entry box and push enter (Pane D in Figure 13).

Detecting the bad anchor:

Using the radio buttons on the bottom left, select **view all anchors**. Scrolling through the patients, you will see the anchors highlighted in red. Scrolling through the first few patients should reveal that "mother" occurs in many contexts other than recent birth, including family history, contact information, etc.

Remove the bad anchor and propose a new one:

We can remove the bad anchor by selecting it and pushing '-.' Now let's add "neonatology" as an anchor. This is a better anchor because it is highly specific. Once again, type it into the anchor entry box, push enter. You can scroll through patients again to confirm that this is a better anchor.

Now next to the suggestions pane, there should be an option 'code,' which will display ICD9 codes in a structured version that can be added to the anchor list with the '+' key. Adding a parent in the hierarchy will add all of the children.

Appendix D

Business Associate Agreement (Executed)

<div style="border:1px solid">

Business Associate Agreement

BUSINESS ASSOCIATE AGREEMENT This Agreement is made effective the ____16th____ day of _January_____, 20_19_____, by and between Century Oak Care Center, Inc., hereinafter referred to as "Covered Entity", and_Dr. Matthew Volansky PT, DPT, MBA_____ hereinafter referred to as "Business Associate" or "Vendor", (individually, a "Party" and collectively, the "Parties").

WITNESSETH: WHEREAS, Sections 261 through 264 of the federal Health Insurance Portability and Accountability Act of 1996, Public Law 104-191, known as "the Administrative Simplification provisions," direct the Department of Health and Human Services to develop standards to protect the security, confidentiality and integrity of health information; and WHEREAS, pursuant to the Administrative Simplification provisions, the Secretary of Health and Human Services has issued regulations modifying 45 CFR Parts 160 and 164 (the "HIPAA Security and Privacy Rule"); and WHEREAS, the Parties wish to enter into or have entered into an arrangement whereby Business Associate will provide certain services to Covered Entity, and, pursuant to such arrangement, Business Associate may be considered a "business associate" of Covered Entity as defined in the HIPAA Security and Privacy Rule (the agreement evidencing such arrangement is entitled Legal Services Agreement, dated _January 16, 2019____, and is hereby referred to as the "Arrangement Agreement"); and WHEREAS, Business Associate may have access to Protected Health Information (as defined below) in fulfilling its responsibilities under such arrangement; THEREFORE, in consideration of the Parties' continuing obligations under the Arrangement Agreement, compliance with the HIPAA Security and Privacy Rule, and for Ten and 00/100s Dollars ($10.00) and other good and valuable consideration, the receipt and sufficiency of which is hereby acknowledged, the Parties agree to the provisions of this Agreement in order to address the requirements of the HIPAA Security and Privacy Rule and to protect the interests of both Parties.

I. DEFINITIONS (a) Except as otherwise defined herein, any and all capitalized terms in this Section shall have the definitions set forth in the HIPAA Security and Privacy Rule. In the event of an inconsistency between the provisions of this Agreement and mandatory provisions of the HIPAA Security and Privacy Rule, as amended, the HIPAA Security and Privacy Rule shall control. Where provisions of this Agreement are different than those mandated in the HIPAA Security and Privacy Rule, but are nonetheless permitted by the HIPAA Security and Privacy Rule, the provisions of this Agreement shall control. (b) The term "Protected Health Information" means individually identifiable health information including, without limitation, all information, data, documentation, and materials, including without limitation, demographic, medical and financial information, that relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and that identifies the individual or with respect to which there is a reasonable basis to believe the information can be used to identify the individual. "Protected Health Information" includes without limitation"Electronic Protected Health

1

</div>

Information" as defined below. (c) The term "Electronic Protected Health Information" means Protected Health Information which is transmitted by Electronic Media (as defined in the HIPAA Security and Privacy Rule) or maintained in Electronic Media. (d) Business Associate acknowledges and agrees that all Protected Health Information that is created or received by Covered Entity and disclosed or made available in any form, including paper record, oral communication, audio recording, and electronic display by Covered Entity or its operating units to Business Associate or is created or received by Business Associate on Covered Entity's behalf shall be subject to this Agreement.

II.   RESPONSIBILITIES OF THE PARTIES WITH RESPECT TO PROTECTIVE HEALTH INFORMATION (a) Business Associate agrees: (i) to use or disclose any Protected Health Information solely: (1) for meeting its obligations as set forth in any agreements between the Parties evidencing their business relationship, or (2) as required by applicable law, rule or regulation, or by accrediting or credentialing organization to whom Covered Entity is required to disclose such information or as otherwise permitted under this Agreement, the Arrangement Agreement (if consistent with this Agreement and the HIPAA Security and Privacy Rule), or the HIPAA Security and Privacy Rule, and (3) as would be permitted by the HIPAA Security and Privacy Rule if such use or disclosure were made by Covered Entity; (ii) at termination of this Agreement, the Arrangement Agreement (or any similar documentation of the business relationship of the Parties), or upon request of Covered Entity, whichever occurs first, if feasible, Business Associate will return or destroy all Protected Health Information received from or created or received by Business Associate on behalf of Covered Entity that Business Associate still maintains in any form and retain no copies of such information, or if such return or destruction is not feasible, Business Associate will extend the protections of this Agreement to the information and limit further uses and disclosures to those purposes that make the return or destruction of the information not feasible; and (iii) to ensure that its agents, including a subcontractor, to whom it provides Protected Health Information received from or created by Business Associate on behalf of Covered Entity, agrees to the same restrictions and conditions that apply to Business Associate with respect to such information, and agrees to implement reasonable and appropriate safeguards to protect any of such information which is Electronic Protected Health Information. In addition, Business Associate agrees to take reasonable steps to ensure that its employees' actions or omissions do not cause Business Associate to breach the terms of this Agreement. (b) Notwithstanding the prohibitions set forth in this Agreement, Business Associate may use and disclose Protected Health Information as follows: (i) if necessary, for the proper management and administration of Business Associate or to carry out the legal responsibilities of Business Associate, provided that as to any such disclosure, the following requirements are met: (A) the disclosure is required by law; or (B) Business Associate obtains reasonable assurances from the person to whom the information is disclosed that it will be held confidentially and used or further disclosed only as required by law or for the purpose for which it was disclosed to the person, and the person notifies Business Associate of any instances of which it is aware in which the confidentiality of the information has been breached; (ii) for data aggregation services, if to be provided by Business Associate for the health care operations of Covered Entity pursuant to any agreements between the Parties

2

evidencing their business relationship. For purposes of this Agreement, data aggregation services means the combining of Protected Health Information by Business Associate with the protected health information received by Business Associate in its capacity as a business associate of another covered entity, to permit data analyses that relate to the health care operations of the respective covered entities. (c) Business Associate will implement appropriate safeguards to prevent use or disclosure of Protected Health Information other than as permitted in this Agreement. Business Associate will implement administrative, physical, and technical safeguards that reasonably and appropriately protect the confidentiality, integrity, and availability of any Electronic Protected Health Information that it creates, receives, maintains, or transmits on behalf of Covered Entity as required by the HIPAA Security and Privacy Rule. (d) Business Associate shall report to Covered Entity any use or disclosure of Protected Health Information which is not in compliance with the terms of this Agreement of which it becomes aware. Business Associate shall report to Covered Entity any Security Incident of which it becomes aware. For purposes of this Agreement, "Security Incident" means the attempted or successful unauthorized access, use, disclosure, modification, or destruction of information or interference with system operations in an information system. In addition, Business Associate agrees to mitigate, to the extent practicable, any harmful effect that is known to Business Associate of a use or disclosure of Protected Health Information by Business Associate in violation of the requirements of this Agreement. (e) Business Associate agrees to make available all records, books, agreements, policies and procedures related to the use and/or disclosure of protected health information to the Secretary of HHS for purposes of determining Covered Entity's compliance with the Privacy Rule. (f) Covered Entity agrees: (i) to obtain any patient consent or authorization that may be required by the Privacy rule or applicable state law prior to furnishing Vendor protected health information pertaining to an individual; (ii) that it will not furnish Vendor protected health information that violates any restrictions on use and/or disclosure as provided for in 45 C.F.R. §164.522 and agreed to by Covered Entity; (iii) to notify Vendor, in writing, of any protected health information in Vendor's possession that Covered Entity seeks to make available to a patient pursuant to 45 C.F.R. §164.524 and agree with Vendor as to the time, manner and form in which Vendor shall provide such access; and (iv) to notify Vendor, in writing, of any amendment(s) to the protected health information in the possession of Vendor that Covered Entity believes are necessary because of its belief that the protected health information that is the subject of the amendment(s) has been or could be relied upon by Vendor or others to the detriment of the individual who is the subject of the protected health information.

III.   AVAILABILITY OF PHI Business Associate agrees to make available Protected Health Information to the extent and in the manner required by Section 164.524 of the HIPAA Security and Privacy Rule. Business Associate agrees to make Protected Health Information available for amendment and incorporate any amendments to Protected Health Information in accordance with the requirements of Section 164.526 of the HIPAA Security and Privacy Rule. In addition, Business Associate agrees to make Protected Health Information available for purposes of accounting of disclosures, as required by Section 164.528 of the HIPAA Security and Privacy Rule.

3

IV.   TERMINATION Notwithstanding anything in this Agreement to the contrary, either party shall have the right to terminate this Agreement and the Arrangement Agreement immediately if that party determines that the other party has violated any material term of this Agreement. Alternatively, either party may choose to provide the other with ten (10) days written notice of the existence of an alleged material breached and afford the breaching party an opportunity to cure. Nonetheless, in the event a mutually agreeable terms cannot be achieved, the breaching party must cure said breach or the agreement shall be terminated. If either party reasonably believes that the other will violate a material term of this Agreement in the future, and where practicable, gives written notice to the other party of such belief, and the other party fails to provide adequate written assurances that it will not breach this agreement, then the party believing that a breach will occur shall have the right to terminate this agreement and the Arrangement Agreement as provided herein.

V.   MISCELLANEOUS (a) Except as expressly stated herein or the HIPAA Security and Privacy Rule, the parties to this Agreement do not intend to create any rights in any third parties. The obligations of Business Associate under this Section shall survive the expiration, termination, or cancellation of this Agreement, the Arrangement Agreement and/or the business relationship of the parties, and shall continue to bind Business Associate, its agents, employees, contractors, successors, and assigns as set forth herein. (b) This Agreement may be amended or modified only in a writing signed by the Parties. No Party may assign its respective rights and obligations under this Agreement without the prior written consent of the other Party. None of the provisions of this Agreement are intended to create, nor will they be deemed to create any relationship between the Parties other than that of independent parties contracting with each other solely for the purposes of effecting the provisions of this Agreement and any other agreements between the Parties evidencing their business relationship. This Agreement will be governed by the laws of the State of Florida. No change, waiver or discharge of any liability or obligation hereunder on any one or more occasions shall be deemed a waiver of performance of any continuing or other obligation, or shall prohibit enforcement of any obligation, on any other occasion. (c) The parties agree that, in the event that any documentation of the arrangement pursuant to which Business Associate provides services to Covered Entity contains provisions relating to the use or disclosure of Protected Health Information which are more restrictive than the provisions of this Agreement, the provisions of the more restrictive documentation will control. The provisions of this Agreement are intended to establish the minimum requirements regarding Business Associate's use and disclosure of Protected Health Information. (d) In the event that any provision of this Agreement is held by a court of competent jurisdiction to be invalid or unenforceable, the remainder of the provisions of this Agreement will remain in full force and effect. In addition, in the event a party believes in good faith that any provision of this Agreement fails to comply with the then-current requirements of the HIPAA Security and Privacy Rule, such party shall notify the other party in writing. For a period of up to thirty days, the parties shall address in good faith such concern and amend the terms of this Agreement, if necessary to bring it into compliance. If, after such thirty-day period, the Agreement fails to comply with the HIPAA Security and Privacy Rule, then either party has the right to terminate upon written notice to the other party. (e) The parties agree to negotiate in good faith mutually acceptable and appropriate amendment(s) to this Agreement to give effect to any amendment to any provision of HIPAA, or its implementing regulations set

4

forth at 45 C.F.R. parts 160 through 164, or any new privacy or security requirements imposed under state or federal law, which materially alters either Party's or both Parties' obligations under this Agreement; provided, however, that if the Parties are unable to agree on mutually acceptable amendment(s) within thirty (30) days of the relevant change of law, either party may terminate this Agreement consistent with section

V. IN WITNESS WHEREOF, the Parties have executed this Agreement as of the day and year written above.

COVERED ENTITY:

Century Oak Care Center, Inc.

Name: Shelly Szarek-Skodny

Address:

7250 Old Oak Blvd.

City/State/Zip:

Middleburg Hts., OH 44130

Telephone: 440-243-7888

Email:

shlely.szarekskodny@centuryoakcarecenter.com

By: _____

01/16/2019

BUSINESS ASSOCIATE:

Name: Dr. Matthew Volansky

Address:

533 Bay Hill Drive

City/State/Zip:

Avon Lake, OH 44012

Telephone: 440-865-0531

Email:

mtv31@shp.rutgers.edu

By: _____

01/16/2019

5
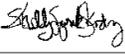
Appendix E

Data Use Agreement.

<div style="border:1px solid #000; padding:1em;">

<div align="center">**DATA USE AGREEMENT**</div>

THIS AGREEMENT is made effective <u>January 16</u>, 20<u>19</u> ("EFFECTIVE DATE") by and between, <u>Dr. Matthew Volansky</u>, ("RECIPIENT") and

- Century Oak Care Center part of the Accord Care Community with its principal place of business at 7250 Old Oak Blvd., Middleburg Heights, OH 44130;

Hereinafter referred to as "**ACC**", and relates to a Limited Data Set which is more particularly described in Attachment A hereof, to be provided by ACC to <u>Dr. Matthew Volansky</u>, of RECIPIENT. The ACC PI is <u>Dr. Matthew Volansky</u>. The purpose of this Agreement is to satisfy certain obligations of ACC under the Health Insurance Portability and Accountability Act of 1996 and its implementing regulations (45 C.F.R. Parts 160-64) ("**HIPAA**") to ensure the integrity and confidentiality of Protected Health Information exchanged in the form of a Limited Data Set.

In consideration of the foregoing and other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, Recipient and ACC agree as follows:

1. <u>Definitions</u>. Capitalized terms used, but not otherwise defined, in this Agreement shall have the meanings given them in HIPAA. For convenience of reference, the definitions of "Individually Identifiable Health Information," "Limited Data Set," and "Protected Health Information" as of the Effective Date are as follows:

1.1 "<u>Individually Identifiable Health Information</u>" means information that is a subset of health information, including demographic information collected from an individual, and (i) is created or received by a healthcare provider, health plan, employer, or health care clearinghouse; and (ii) relates to the past, present, or future physical or mental health or condition of an individual; the provision of healthcare to an individual; or the past, present, or future payment for the provision of health care to an individual; and (a) that identifies the individual, or (b) with respect to which there is a reasonable basis to believe the information can be used to identify the individual.

1.2 "<u>Limited Data Set</u>" means Protected Health Information that excludes the following direct identifiers of the individual or of relatives, employers, or household members of the individual: (i) Names; (ii) Postal address information, other than town or city, State, and zip code; (iii) Telephone numbers; (iv) Fax numbers; (v) Electronic mail addresses; (vi) Social security numbers; (vii) Medical record numbers; (viii) Health plan beneficiary numbers; (ix) Account numbers; (x) Certificate/license numbers; (xi) Vehicle identifiers and serial numbers, including license plate numbers; (xii) Device identifiers and serial numbers; (xiii) Web Universal Resource Locators (URLs); (xiv) Internet Protocol (IP) address numbers; (xv) Biometric identifiers, including finger and voice prints; and (xvi) Full face photographic images and any comparable images.

1.3 "<u>Protected Health Information</u>" means Individually Identifiable Health Information that Recipient receives from ACC or from a business associate of ACC or which Recipient creates for ACC which is transmitted or maintained in any form or medium. "Protected Health Information" shall not include education records covered by the Family Educational Right and Privacy Act, as

</div>

amended, 20 U.S.C. §1232g, or records described in 20 U.S.C. §1232g (a)(4)(B)(iv), or employment records held by ACC in its role as employer.

2. **Applicability of Terms; Conflicts.** This Agreement applies to the Limited Data Set as described in Attachment A hereto.

3. **Obligations and Activities of Recipient**

3.1 **Non-disclosure:** Recipient will not use or disclose the Limited Data Set other than as permitted or required by this Agreement or as Required By Law or as otherwise authorized by ACC.

3.2 **Safeguards:** Recipient will use appropriate safeguards to prevent use or disclosure of the Limited Data Set other than as provided for by this Agreement. Recipient will develop, implement, maintain and use appropriate administrative, technical and physical safeguards to preserve the integrity and confidentiality of and to prevent non-permitted or violating use or disclosure of the Limited Data Set which is transmitted electronically. Recipient will document and keep these safeguards current.

3.3 **Mitigation:** Recipient will mitigate, to the extent practicable, any harmful effect that is known to Recipient of a use or disclosure of the Limited Data Set by Recipient in violation of the requirements of this Agreement.

3.4 **Reporting:** Recipient will report to the Privacy Officer of ACC, in writing, any use and/or disclosure of the Limited Data Set that is not permitted or required by this Agreement of which Recipient becomes aware. Such report shall be made as soon as reasonably possible but in no event more than five (5) business days after discovery by Recipient of such unauthorized use or disclosure. This reporting obligation shall include breaches by Recipient, its employees, subcontractors and/or agents. Each such report of a breach will, to the extent possible: (i) identify the nature of the non-permitted or violating use or disclosure; (ii) identify the Limited Data Set used or disclosed; (iii) identify who received the non-permitted or violating use or disclosure; (iv) identify what corrective action Recipient took or will take to prevent further non-permitted or violating uses or disclosures; (v) identify what Recipient did or will do to mitigate any deleterious effect of the non-permitted or violating use or disclosure; and (vi) provide such other information as ACC may reasonably request.

3.5 **Agents and Subcontractors:** Recipient will ensure that any agent, including a subcontractor, to whom it provides the Limited Data Set received from, or created or received by Recipient on behalf of, ACC agrees to the same restrictions and conditions that apply through this Agreement to Recipient with respect to such information.

3.6 **Identification and Contact of Individuals:** Recipient will not identify or attempt to identify the individuals whose Protected Health Information appears in the Limited Data Set. Recipient will not contact or attempt to contact the individuals whose Protected Health Information appears in the Limited Data Set.

4. **Permitted Uses and Disclosures by Recipient.**

    4.1 **Health Care Operations, Public Health and Research.** Except as otherwise limited in this Agreement, Recipient may use or disclose the Limited Data Set only for purposes of research, public health or Health Care Operations.

5. **Term and Termination**

    5.1 **Term.** The term of this Agreement shall commence as of the Effective Date, and shall terminate when all of the Limited Data Set provided by ACC to Recipient, or created or received by Recipient on behalf of ACC, is destroyed or returned to ACC, or, if it is infeasible to return or destroy the Limited Data Set, protections are extended to such Limited Data Set in accordance with the provisions of this Section 5.

    5.2 **Termination for Cause.** Upon ACC's reasonable determination that Recipient has breached a material term of this Agreement, ACC shall be entitled to do any one or more of the following:

        Give Recipient written notice of the existence of such breach and give Recipient an opportunity to cure upon mutually agreeable terms. If Recipient does not cure the breach or end the violation according to such terms, or if ACC and Recipient are unable to agree upon such terms, ACC may immediately terminate this Agreement. Simultaneously, ACC may immediately stop all further disclosures of the Limited Data Set to Recipient.

    5.3 **Effect of Termination.** Upon receipt of written demand from ACC, Recipient agrees to immediately return or destroy, except to the extent infeasible, all of the Limited Data Set demanded by ACC, including all such Limited Data Set which Recipient has disclosed to its employees, subcontractors and/or agents. Destruction shall include destruction of all copies including backup tapes and other electronic backup medium. In the event the return or destruction of some or all such Limited Data Set is infeasible, the Limited Data Set not returned or destroyed pursuant to this paragraph shall be used or disclosed only for those purposes that make return or destruction infeasible.

    5.4 **Continuing Privacy Obligations.** Recipient's obligation to protect the privacy of the Limited Data Set is continuous and survives any termination, cancellation, expiration, or other conclusion of this Agreement with respect to any portion of the Limited Data Set Recipient maintains after such termination, cancellation, expiration or other conclusion of this Agreement.

6. **Notices.** All notices pursuant to this Agreement must be given in writing and shall be effective when received if hand-delivered or upon dispatch if sent by reputable overnight delivery service, or U.S. Mail to the appropriate address as set forth on the last page of this Agreement or via email.

7. **Miscellaneous.** Recipient and ACC agree that individuals whose Protected Health Information appears in a Limited Data Set are not third-party beneficiaries of this Agreement. In the event that any provision of this Agreement violates any applicable statute, ordinance or rule of law in any jurisdiction that governs this Agreement, such provision shall be ineffective to the extent of such violation without invalidating any other provision of this Agreement. This Agreement may not be amended, altered or modified except by written agreement signed by Recipient and ACC. No provision of this Agreement may be waived except by an agreement in writing signed by the waiving

party. A waiver of any term or provision shall not be construed as a waiver of any other term or provision. Nothing in Section 3 of this Agreement shall be deemed a waiver of any legally-recognized claim of privilege available to Recipient. The persons signing below have the right and authority to execute this Agreement for their respective entities and no further approvals are necessary to create a binding agreement. Neither ACC nor Recipient shall use the names or trademarks of the other party or of any of the respective party's affiliated entities in any advertising, publicity, endorsement, or promotion unless prior written consent has been obtained for the particular use contemplated. All references herein to specific statutes, codes or regulations shall be deemed to be references to those statutes, codes or regulations as may be amended from time to time.

IN WITNESS WHEREOF, the parties have executed this Agreement as of the day and year referenced above

Century Oak Care Center, Inc.

Name: Shelly Szarek-Skodny   Name: Dr. Matthew Volansky

Address: Address:

7250 Old Oak Blvd.   533 Bay Hill Drive

City/State/Zip: City/State/Zip:

Middleburg Hts., OH 44130   Avon Lake, OH 44012

Telephone: 440-243-7888   Telephone: 440-865-0531

Email: Email:

shlely.szarekskodny@centuryoakcarecenter.com   mtv31@shp.rutgers.edu
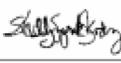
By: _____   By: _____

01/16/2019   01/16/2019

Appendix F

IRB Approval communication

Fw: IRB Approval #353

Volansky, Matthew T
Sun 4/28/2019 2:49 PM
To: Dr. Matt Volansky <mtv31@shp.rutgers.edu>

From: IRB
Sent: Sunday, April 28, 2019 2:19 PM
To: Volansky, Matthew T
Subject: IRB Approval #353

IRB Approval #353

Dear Matt,

The Institutional Review Board (IRB) has reviewed the above-listed research. According to the information you provided, this proposal meets the IRB requirements and has been approved. You are authorized to begin the research under the number assigned above.

Please note that any complaints received by you from research participants must be reported to the IRB. The IRB will determine whether the complaint meets the definition of an: unanticipated problem or non-compliance of the research protocol.

It is understood this project will be conducted in full accordance with all applicable sections of the most current version of the IRB Guidelines. It is also understood that the IRB will be immediately notified of any changes that may affect the review status of your research project.

Sincerely,

Ron Mendel, Ph.D.
Chair, University of Mount Union IRB
Department of Human Performance & Sport Business
University of Mount Union
Alliance, OH 44601

Appendix G

Expert Physical Therapist Anchor Variable Selection Survey

Semi-Supervised Electronic Phenotyping Via the Anchor and Learn Framework with Physical Therapy Emphasis.

1. Expert Anchor Term Identification

Your task is to appraise each of the words or phrases by asking the following question:

"How confident am I that if present in the documentation, this word or phrase is likely to represent the presence of a suspected or confirmed VTE?"

Words and phrases should be examined for their utility in predicting the risk for the patient in contracting, having or exacerbating a VTE diagnosis. You are asked to use your knowledge of not only explicit events which lead to VTE disease, but also for the subtle physical therapy based factors which allow for the increased risk for VTE to develop. In as such, these factors may not be concrete and appear vague to the untrained clinician. Such vague terms may hold the same weight as the terms "dvt" and "pe" themselves. Individual words which are part of a word pair should be treated individually. For example the individual words "chest" and "pain" individually may have little value in identifying VTE. However the combination of the words "chest_pain" may have a much higher likelihood of VTE presence.

Please rank each of the word or phrases for their ability to identify the presence of either undiagnosed, acute, or chronic VTE. A 5 point Likert scale is utilized with the following rank definitions:

1 = extremely unlikely
2 = unlikely
3 = neutral
4 = likely
5 = extremely likely

Please select only one likelihood response for each word / phrase which are presented in alphabetical order.

OK

* 1. "How confident am I that if present in the documentation, this word or phrase is likely to represent a the presence of a suspected or confirmed VTE?"

	extremely unlikely	unlikely	neutral	likely	extremely likely
abnormality	✔				
activity_tolerance				✔	
acute			✔		
ADLs			✔		
air	✔				
ambulation			✔		
anemia			✔		
ankle_foot				✔	
aortic			✔		
arthroplasty				✔	
bed				✔	
bed_mobility				✔	
bed_to_wc_wheelchair				✔	
blood			✔		
board		✔			
boot				✔	
boot_afo				✔	

* 2. "How confident am I that if present in the documentation, this word or phrase is likely to represent a the presence of a suspected or confirmed VTE?"

	extremely unlikely	unlikely	neutral	likely	extremely likely
ca_cancer				✔	
CABG				✔	
CAD				✔	
cardiac			✔		
catheter			✔		
chest_pain			✔		
clot					✔
co_complainsof			✔		
co_SOB				✔	
compression				✔	
CT_ComputedTomography				✔	
do_diagnosis		✔			
debility		✔			
decrd		✔			
deep				✔	
deep_vessels				✔	
develop		✔			
difficulty		✔			
difficulty_walking				✔	
dorsiflex				✔	
DVT					✔
dyspepsia		✔			
dyspnea				✔	

* 3. "How confident am I that **if present** in the documentation, this word or phrase is likely to represent a the presence of a suspected or confirmed VTE?"

	extremely unlikely	unlikely	neutral	likely	extremely likely
edema	○	○	○	✔	○
embolism	○	○	○	○	✔
endurance	○	○	○	✔	○
eob_edgeofbed	○	○	✔	○	○
er_emergencyroom	○	○	✔	○	○
fall	○	○	○	✔	○
fall_withfx	○	○	○	✔	○
fall_risk	○	○	○	✔	○
femoral	○	○	✔	○	○
femoral_vein	○	○	○	✔	○
filter	○	○	○	✔	○
foot	○	○	✔	○	○
fracture	○	○	○	✔	○
functional_activity	○	○	○	✔	○
groin	○	○	✔	○	○
hematoma	○	○	✔	○	○
hemorrhage	○	○	✔	○	○
hip	○	○	✔	○	○
history	○	○	✔	○	○
home	○	○	✔	○	○
hospital	○	○	○	✔	○
hoyer_lift	○	○	○	✔	○
hypercoagul	○	○	○	○	✔

* 4. "How confident am I that **if present** in the documentation, this word or phrase is likely to represent a the presence of a suspected or confirmed VTE?"

	extremely unlikely	unlikely	neutral	likely	extremely likely
implant	○	○	○	✔	○
improved_strength	○	○	✔	○	○
infection	○	○	○	✔	○
inferior	○	○	✔	○	○
infiltrate	○	○	✔	○	○
inoutbed	○	○	✔	○	○
insufficiency	○	○	✔	○	○
ivc_filter	○	○	○	✔	○
knee	○	○	✔	○	○
LE	○	○	○	✔	○
leg	○	○	○	✔	○
lethargic	○	○	✔	○	○
limit	○	○	✔	○	○
LLE	○	○	○	✔	○
lobe	○	○	○	✔	○
LTACH	○	○	○	✔	○
lymphoma	○	○	○	✔	○
mass	○	○	✔	○	○
max_assist	○	○	○	✔	○
mod_max_assist	○	○	○	✔	○
movement	○	○	○	✔	○

	extremely unlikely	unlikely	neutral	likely	extremely likely
neck	○	○	✔	○	○
nonhealing	○	○	✔	○	○
nursing	○	○	✔	○	○
obesity	○	○	○	✔	○
obstructive	○	○	✔	○	○
orthocare	○	○	✔	○	○
pain	○	○	✔	○	○
pain_joint	○	○	✔	○	○
PE	○	○	○	○	✔
pe_pulmonaryembolism	○	○	○	○	✔
pelvis	○	○	✔	○	○
positive	○	○	✔	○	○
post	○	○	✔	○	○
pre	○	○	✔	○	○
presented	○	○	✔	○	○
PROM	○	○	✔	○	○
pulmonary	○	○	○	✔	○
pulmonary_emboli	○	○	○	✔	○
replacement	○	○	○	✔	○
rest	○	○	✔	○	○
RLE	○	○	○	✔	○
ROM	○	○	✔	○	○
room	○	○	✔	○	○

	extremely unlikely	unlikely	neutral	likely	extremely likely
sci_spinalcordinjury	○	○	○	✔	○
severe	○	○	✔	○	○
shin	○	○	✔	○	○
sit	○	○	✔	○	○
sitting	○	○	✔	○	○
skin_integrity	○	○	✔	○	○
slide	○	○	✔	○	○
SOB	○	○	○	✔	○
sock	○	○	✔	○	○
spo2	○	○	○	✔	○
standing_pivot	○	○	✔	○	○
superior	○	○	✔	○	○
surgery	○	○	○	✔	○
swell	○	○	○	✔	○
tachycardia	○	○	○	✔	○
thigh	○	○	✔	○	○
thrombosis	○	○	○	✔	○
tibi	○	○	✔	○	○
time	○	○	✔	○	○
TKA	○	○	○	✔	○
traumatic	○	○	○	✔	○
TTWB	○	○	○	✔	○

* 7. "How confident am I that if present in the documentation, this word or phrase is likely to represent a the presence of a suspected or confirmed VTE?"

	extremely unlikely	unlikely	neutral	likely	extremely likely
ulcer	○	○	✔	○	○
vein	○	○	✔	○	○
venous	○	○	○	✔	○
venous_embolism	○	○	○	✔	○
vessel	○	○	✔	○	○
wc_wheelchair	○	○	✔	○	○
weakness	○	○	✔	○	○
well	○	○	✔	○	○
wound_vac	○	○	✔	○	○
wrap	○	○	✔	○	○

DONE

TABLE 20. Anchor Token Maximum Frequency Analysis Expert Physical Therapist.

Table	Anchor Token	Rank Order		Frequency Count		Percentage	
		First	Top Five	First	Top Five	First	Top Five
Expert Physical Therapist	activity_tolerance	14	14	0	3	0%	0.9%
	adl	13	13	0	9	0%	2.7%
	ambulation	15	15	0	0	0%	0%
	arthroplasty *	Removed					
	bed_mobility	16	16	0	0	0%	0.0%
	boot	7	1	6	41	6.7%	12.2%
	cabg	3	7	10	23	11.2%	6.8%
	cad	4	4	9	32	10.1%	9.5%
	clot *	Removed					
	co_sob *	Removed					
	deep_vessels *	Removed					
	develop	2	3	14	34	15.7%	10.1%
	dvt	10	11	3	20	3.4%	6.0%
	edema	12	9	0	23	0%	6.8%
	filter	1	6	17	23	19.1%	6.8%
	hypercoagul *	Removed					
	lobe	11	12	3	18	3.4%	5.4%
	movement	8	2	6	35	6.7%	10.4%
	pe	9	8	3	23	3.4%	6.8%
	pulmonary	5	5	9	31	10.1%	9.2%
	shin	6	10	9	21	10.1%	6.3%
	thrombosis *	Removed					
	vein *	Removed					
	venous_embolism *	Removed					
	vessel *	Removed					
				89	336	100%	100%

* Removed from model secondary to sparseness of data causing prediction error

TABLE 21. Anchor Token Maximum Frequency Analysis Academy of Acute Care Physical Therapy CPG.

Table	Anchor Token	Rank Order		Frequency Count		Percentage	
		First	Top Five	First	Top Five	First	Top Five
Academy Acute Care Physical Therapy (CPG)	ambulate	13	12	2	26	3.0%	4.1%
	cancer	3	6	3	41	4.5%	6.4%
	edema	12	13	3	25	4.5%	3.9%
	embolus *	Removed					
	inflammatory *	Removed					
	mechanical	9	11	5	34	7.6%	5.3%
	compression	5	4	11	44	16.7%	6.9%
	surgery	15	15	0	17	0.0%	2.7%
	calf	2	3	2	57	3.0%	8.9%
	family	10	10	4	36	6.1%	5.6%
	swell	7	5	9	43	13.6%	6.7%
	immobilization *	Removed					
	anticoagul	8	7	6	40	9.1%	6.3%
	extremity *	Removed					
	hospital	16	18	0	0	0%	0%
	leg	17	16	0	6	0%	0.9%
	pressure	14	8	2	40	3.0%	6.3%
	ultrasound *	Removed					
	heparin *	Removed					
	dvt	1	1	1	100	1.5%	15.7%
	prophylaxis	4	2	4	66	6.1%	10.3%
	pulmonary	11	14	4	22	6.1%	3.4%
	fall	18	17	0	1	0%	0.2%
	vein	6	9	10	40	15.2%	6.3%
	wells *	Removed					
				66	638	100%	100%

Removed from model secondary to sparseness of data causing prediction error

TABLE 22. Anchor Token Maximum Frequency Analysis Clinician Documentation.

Table	Anchor Token	Rank Order		Frequency Count		Percentage	
		First	**Top Five**	**First**	**Top Five**	**First**	**Top Five**
Clinician Documentation	rom	8	9	6	54	2.2%	4.2%
	home	15	19	3	33	1.1%	2.6%
	le	24	24	0	23	0%	1.8%
	sit	7	10	7	49	2.5%	3.8%
	acute	12	7	5	63	1.8%	4.9%
	wc_wheelchair	4	5	12	64	4.4%	4.9%
	ambulation	22	13	1	44	0.4%	3.4%
	time	19	12	2	44	0.7%	3.4%
	anemia	9	15	6	40	2.2%	3.1%
	lle	20	18	2	36	0.7%	2.8%
	functional_activity	13	17	4	37	1.5%	2.9%
	activity_tolerance	16	23	3	23	1.1%	1.8%
	weakness	17	3	3	73	1.1%	5.6%
	bed_mobility	14	14	4	41	1.5%	3.2%
	difficulty	23	20	1	32	0.4%	2.5%
	endurance	6	4	9	71	3.3%	5.5%
	fall_risk	10	22	6	29	2.2%	2.2%
	foot	5	8	12	60	4.4%	4.6%
	hospital	18	11	3	47	1.1%	3.6%
	fall	11	21	6	30	2.2%	2.3%
	filter	1	1	141	206	51.3%	15.9%
	pain	25	25	0	13	0%	1.0%
	leg	2	6	20	63	7.3%	4.9%
	ulcer	3	2	17	80	6.2%	6.2%
	mobility	21	16	2	39	0.7%	3.0%
				275	1294	100%	100%

* Removed from model secondary to sparseness of data causing prediction error

## Academy of Acute Care Physical Therapy Clinical Practice Guideline

The Role of Physical Therapists in the Management of Individuals at Risk for or Diagnosed with Venous Thromboembolism - An Evidence-Based Clinical Practice Guideline.

As accessed at https://academic.oup.com/ptj/article/96/2/143/2686356/

Figure 14. Algorithm for screening for risk of venous thromboembolism (VTE).
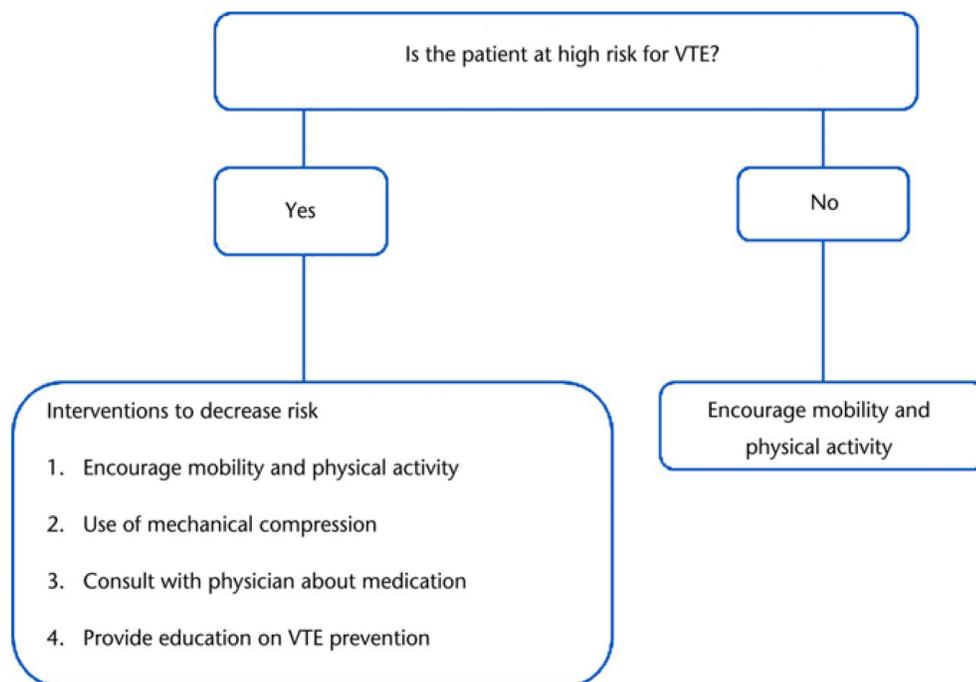
Figure 15. Algorithm for determining likelihood of a lower extremity deep vein thrombosis (LE DVT).
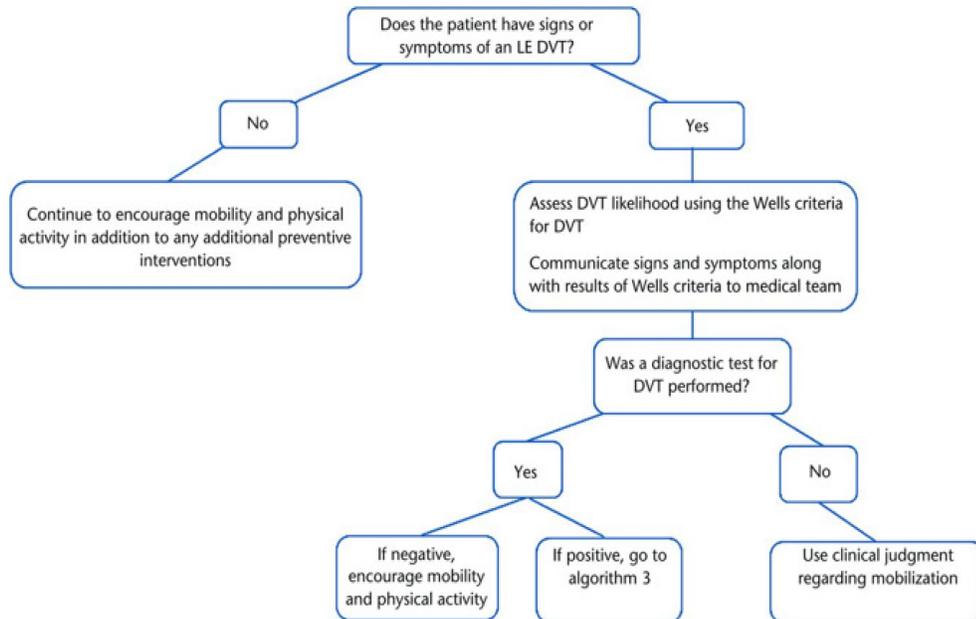
OXFORD
UNIVERSITY PRESS

Figure 16. Algorithm for mobilizing patients with known lower extremity deep vein thrombosis (LE DVT).

OXFORD
UNIVERSITY PRESS