University of Vermont
UVM ScholarWorks

Graduate College Dissertations and Theses | Dissertations and Theses

2022

Building a Learning Healthcare System: a path to optimizing big health data to inform clinical care decisions

Danne Charlotte Emily Elbers
University of Vermont

Follow this and additional works at: https://scholarworks.uvm.edu/graddis

 Part of the Bioinformatics Commons, Computer Sciences Commons, and the Medical Sciences Commons

Recommended Citation

Elbers, Danne Charlotte Emily, "Building a Learning Healthcare System: a path to optimizing big health data to inform clinical care decisions" (2022). *Graduate College Dissertations and Theses*. 1550.
https://scholarworks.uvm.edu/graddis/1550

This Dissertation is brought to you for free and open access by the Dissertations and Theses at UVM ScholarWorks. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of UVM ScholarWorks. For more information, please contact scholarworks@uvm.edu.

Building a Learning Healthcare System: a path to optimizing big health data to inform clinical care decisions

A Dissertation Presented

by

Danne Charlotte Emily Pieneman - Elbers

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Complex Systems and Data Science

May, 2022

Defense Date: March 23rd, 2022
Dissertation Examination Committee:

Peter Dodds, Ph.D., Advisor
Chris Danforth, Ph.D., Advisor
Sarah A. Nowak, Ph.D., Chairperson
Robert E. Gramling, M.D., D.Sc.
Safwan Wshah, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

The explosive growth of data and computing power of the last decades has had large impacts on a myriad of domains, not in the least on one of society's most complex systems: healthcare. In this work, a version of the resulting Learning Healthcare System (LHS) is explored and elements of it have been implemented and are in use at the Department of Veterans' Affairs today. After an overview of what a LHS is and what it could be once executed in its full form, the chapters will describe in detail some of the individual elements and how they address cogs of the LHS' cyclic system. A data repository and clinical knowledge base to facilitate the LHS, called the Precision Oncology Data Repository (PODR), will be highlighted, as will two applications: one addressing clinical trial enrollment at point of care, and another synchronization data back into the clinics and hospitals at the (on-going) time of the COVID-19 epidemic. Both these applications are heavily utilizing the large Electronic Health Record real-world data to generate actionable knowledge while applying advanced analytics. Lastly, this thesis presents a methodology for re-calibrating and validating sentiment analysis of EHR clinical notes to facilitate a near real-time pulse of the interaction between patients, providers and the hospital, with the goal of delivering new insights and allowing for iterative adaption based on measurable performance.

Citations

Material from this dissertation has been published in the following form:

Dhond, R. & Elbers, D. C., Majahalme, N., Dipietro, S., Goryachev, S., Acher, R., Leatherman, S., Anglin-Foote, T., Liu, Q., Su, S., Seerapu, R., Hall, R., Ferguson, R., Brophy, M. T., Ferraro, J., Duvall S. L. and Do, N. V.. (2021). ProjectFlow: a configurable workflow management application for point of care research. *JAMIA Open*, 4(3), ooab074.

Elbers, D. C. & Fillmore, N. R., Sung, F., Ganas, S., Prokhorenkov, A., Meyer, C., Hall, R. B., Ajjarapu, S. J., Chen, D. C., Meng, F., Grossman, R. L., Brophy, M. T. and Do, N. V.. (2020). The Veterans Affairs Precision Oncology Data Repository, a Clinical, Genomic, and Imaging Research Database. Patterns, 1(6), 100083.

Fillmore, N. R. & Elbers, D. C., La, J., Feldman, T. C., Sung, F., Hall, R. B., Nguyen, V., Link, N., Zwolinski, R., Dipietro, S., Miller, S. J., Aleksanyan, A., Goryachev, S., Corcoran, P., Bergstrom, S. J., Parenteau, M. A., Sprague, R. S., Thornton, D. J., Driver, J. A., Strymish, J. M., Evans, S., Colonna, B., Brophy, M. T. and Do, N. V.. (2020). An application to support COVID-19 occupational health and patient tracking at a Veterans Affairs medical center. Journal of the American Medical Informatics Association, 27(11), 1716-1720.

Please note that work listed above is of such size that a sole first author is not a fair representation. Sequence of occurrence has been determined by a coin-flip (Dhond & Elbers), or a 'tit-for-tat' strategy (Elbers & Fillmore, Fillmore & Elbers).

Material from this dissertation has been submitted for publication to PLOS ONE on (04, 15, 2022) in the following form:

Elbers, D. C., La, J., Minot, J., Gramling, R. E., Brophy, M. T., Do, N. V., Fillmore, N., Dodds, P., Danforth, C.. (2022). Sentiment analysis of medical record notes for lung cancer patients in the Department of Veterans Affairs.

In dedication to my husband,

who encourages me to be more and more inquisitive each day and,

when answers lack, shows that courage remains.

To our amazing children,

who taught me the real meaning of focus.

Acknowledgements

There are many people I'd like to express my gratitude to and without whose support, directly or indirectly, this thesis would not exist. First, I'd like to thank my advisors Peter Dodds and Chris Danforth for their unwavering guidance throughout these last few years and their encouragement to combine research in healthcare with the complex systems and data science field. It truly helped me create a deeper understanding and vision leading to this final dissertation.

A heartfelt thank you as well to my committee, Robert Gramling, Safwan Wshah and Sara Nowak for joining into the latter stages of this research adventure and offering encouragement and feedback. I'd like to extend my thanks as well to the Computational Storylab and all the people that I've met and inspired me throughout the years; Josh Minot, Mikaela Fudolig, Thayer Alshaabi, David Dewhurst, Jane Adams and many others. Although this time focused around social distancing made it hard to form close connections, I've always enjoyed and appreciated your collaborative, critical and creative mind set.

My deep appreciation goes to all my colleagues at the Department of Veterans Affairs that inspire me every day to improve and continue working towards the best healthcare we can provide. Svitlana Dipietro, Samuel Ajjarapu, Jennifer La, Ramin Pourali, Alex Shayan, Frank Meng, Feng-Chi (Robert) Sung, Nilla Majahalme and Rupali (Polly) Dhond and all others, this journey would not have been possible without you. I especially would like to thank Mary Brophy, Nhan Do and Nathanael Fillmore for their believe in me and their ever-present support.

A myriad of warm thank you's to my parents, Caroline and Kier Elbers and my broth-

ers, for inspiration, countless hours of babysitting, keeping things light and teaching me how to explain and communicate my work. Warm gratitude goes out as well to my parents-in-law, for all their support, love and understanding. Lastly, no words - trust me, I've tried - can express my gratitude to my husband, Jerry Pieneman, for standing by my side these last years, I would not be here without you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Historic view of the Electronic Healthcare Record System

The evolution in medical record keeping through creating an electronic health record system (EHR) started near the end of the '60s in the United States, with the Massachusetts General Hospital (MGH) being the first adopter utilizing the Computer Stored Ambulatory Record (COSTAR) system, which was developed in collaboration with Harvard (Barnett et al., 1979). MGH's implementation was quickly followed by the Department of Veterans Affairs' (VA) adoption and nationwide roll out of the VA health information system and technology architecture, commonly known as VistA, during the '70s till mid '80s (Atherton, 2011)(Brown, 2003)(Olsen, Aisner, & McGinnis, 2007).

Originally developed for billing purposes, improving the standards of clinical care through better record keeping and an increase in quality assurance and improvement capabilities (Hersh, 2007)(I. o. M., 2003)(Bates et al., 1999), the roll out of EHRs signified the start of a multitude of new health research areas. It didn't take long for the first research projects were proposed taking advantage of this new wealth of data, and secondary data use became adopted within medical research. Since then, EHRs all over the world have been initiated, implemented in clinics and hospitals, and continue to grown in size and types of data they support (Wen, Ho, Jian, Li, & Hsu, 2007)(Casey, Schwartz, Stewart, & Adler, 2016). Today, at the VA alone, the EHR data systems have expanded to supporting over 1.5 petabyte of data.

1.1.2 Healthcare Data Science and Complex Systems Research

In addition to its ongoing growth in volume, healthcare data has grown substantially in complexity due to its utilization, ever-expanding treatment pathways, and addition of new techniques generating new data types (Greenhalgh & Papoutsi, 2018)(Burton, Elliott, Cochran, & Love, 2018). Nowadays a patient encounter with a given medical center can generate a large set of diverse data elements for example; clinical data, imaging scans (MRI, PT, CT), all sorts of genomic sequencing results, pathology tissue and digital pathology slide scans. Data types range from: standard elements such as structured and unstructured text, date-time stamps, Clinical Modification Diagnoses (ICD-9-CM and ICD-10-CM) and prescription (Rx) codes housed in large data warehouses, to Digital Imaging and Communication in Medicine (DICOMs)(Mustra, Delac, & Grgic, 2008), FASTQs, PDFs, VCFs and BAMs, and the list doesn't end there.

It is not surprising that researchers jumped in enthusiastically to make, secondary or primary, use of this plethora of healthcare data. For example, in population research and epidemiology, large-scale studies have taken advantage of secondary data use, creating research cohorts that easily range over >10.000 patients (Casey et al., 2016). While in the Natural Language Processing domain of Machine Learning work has been done building algorithms to extract information from clinical notes (Sheikhalishahi et al., 2019)(Alba et al., 2021)(Meng, Morioka, & Elbers, 2019). Machine learning approaches and data science research in healthcare has taken off in general (Ghassemi et al., 2020)(Ben-Israel et al., 2020), even so much so that the VA recently started

their own Artificial Intelligence Institute (HealthITAnalytics, 2019).

1.1.3 Gap between Research and Clinical Implementation

Idealistically, the goal of healthcare related research is to improve patient outcomes and the quality of care patients receive. Unfortunately, it can take years for any given research outcome to make it back into the clinic (Morris, Wooding, & Grant, 2011)(Budrionis & Bellika, 2016). This is often called the 'bench to bedside' gap and exists both in the traditional clinical trial type research as it does in the newer translational and, machine learning or data science research approaches (Ben-Israel et al., 2020)(van der Laan & Boenink, 2015). The gap between research outcomes and clinical implementation and/or adoption can be attributed to several causes. In the domain of traditional clinical trial type research, it takes time to validate new findings and embed them into current clinical standards and practice. Whereas in machine learning and data science, transparency of the algorithms and delivery of them back into the clinic are often named as obstacles (Ben-Israel et al., 2020)(Hakkoum, Abnane, & Idri, 2022).

1.2 The Learning Healthcare System

1.2.1 Definition

In 2006, the Institute of Medicine held a round-table conference to address the issue outlined above and introduced and defined the Learning Healthcare System (LHS) as: "A learning healthcare system is one that is designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care". (Olsen et al., 2007) Unlike research efforts with traditional clinical trials where the emphasis is on the population under ideal conditions, LHS's attention is on treatment and interventions that are optimized for the patients under real world conditions.

1.2.2 Components

Multiple models for implementation of a LHS have been proposed to meet the mandate issued by the Institute of Medicine to transform healthcare systems. Inherent to all of them are three essential infrastructure-related activities supporting a learning cycle. These are

(A) the creation of clinical knowledge bases to integrate and manage a growing volume and variety of data,

(B) the generation of actionable knowledge using real-world evidence and advanced analytics and,

(C) the delivery and application of these newly discovered insights (knowledge) to

improve patient care as well as the iterative adaptation based on performance.

Each of these three core activities encompasses multiple informatics approaches and technological challenges. In the learning cycle described by Friedman et al. (Friedman, Rubin, & Sullivan, 2017) there is often a disconnect or gap between the research activities that generate the knowledge from data and the clinical operational activities of implementing knowledge to improve performance, as outlined above. This work highlights three publications on multiple projects that are underway to inform not only the technology but the people, policy, and processes necessary to integrate both research and clinical activities for continuous iterations of learning.

Since 2007, numerous LHS models have been described, each encompassing and building on the three core tasks described above. In their LHS review, McLachlan et al. (McLachlan et al., 2018) noted a general lack of consistency in the description of most LHS efforts and proposed a taxonomy which decomposes the three core tasks into nine primary archetypes of which six rely on cohort identification. The LHS taxonomy has been further advanced by Wouters et al. (Wouters, van der Graaf, Voest, & Bredenoord, 2020) who described four LHS implementation models (Optimization LHS, Comprehensive Data LHS, Real time LHS, Full LHS). A Full LHS implementation model has workflow components that optimize the analysis of observational data, deliver real time actionable knowledge through clinical decision support, and inform treatment decision through embedded clinical trials. The following chapter (chapter 2) describes the development and implementation of an application (ProjectFlow), that successfully embeds clinical trials into the clinical care workflow to facilitate national point-of-care research (Dhond et al., 2021).

Subsequently in chapter 3 the generation of a knowledge repository, the Precision Oncology Data Repository (PODR)(Elbers et al., 2020), is addressed fulfilling element (A) of a LHS described by Friedman et al. Both of these chapters relate to an LHS system specifically in oncology, a healthcare domain highlighted as very well suited for this purpose (Do et al., 2019). The proposed bigger, and partially implemented, picture for a LHS in oncology at the VA can be seen in figure 1.1. In chapter 5, research is described re-calibrating and applying a social-science sentiment measurement instrument, the Hedonometer, (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011) to clinical oncology notes, in order to better understand the interaction between patient, provider and healthcare system (Elbers et al., 2022). This allows for the execution of the 'Generated Knowledge & Algorithms' loop in figure 1.1 and item (B) according to Friedman's LHS description. Friedman's item (C), can be partially found in chapter 2 mentioned above, but is also addressed in chapter 4 through an application developed to assist the VA Boston during the start of the COVID-19 epidemic. This framework has subsequently been reused in the oncology LHS and deployed as a application platform housing a clinical trial matching and tumor-board app (manuscript in development). This last piece completes the full circle of a LHS as described above, though its implementation is lean at the moment. Over the upcoming years, the LHS will benefit from the addition of many components and constant tweaking and tailoring to ensure a continuous optimization of data, enabling clinical care decisions.

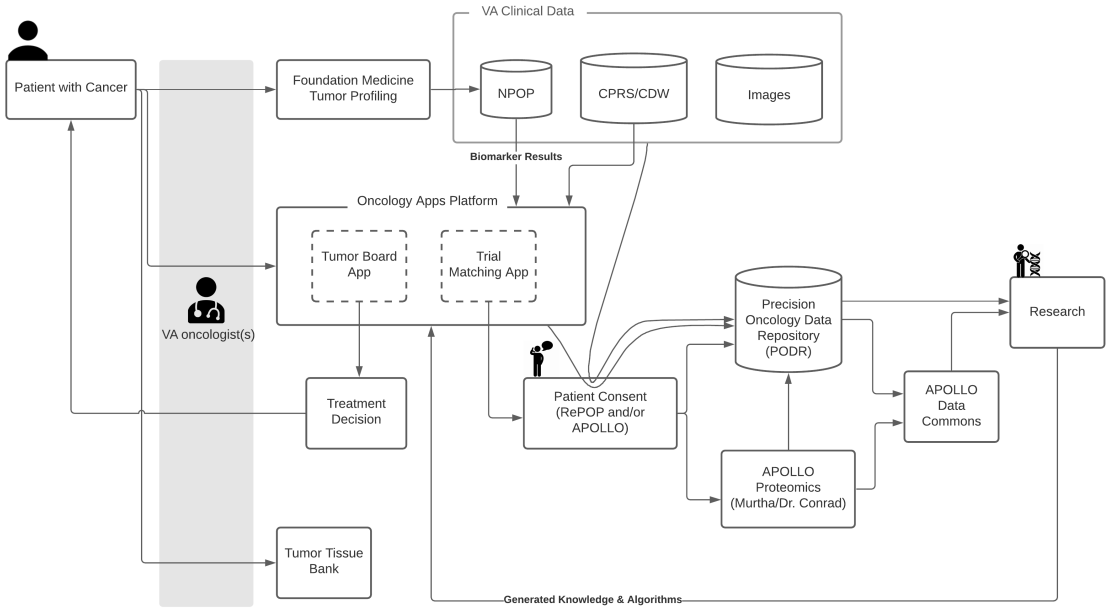


Figure 1.1: Precision Oncology Learning Healthcare System; flowchart diagram shows the partially implemented, partially planned lay-out of a LHS in oncology at the department of Veterans Affairs.

Chapter 2

ProjectFlow: a configurable workflow management application for point of care research

This Chapter is derived from Dhond & Elbers et al. (Dhond et al., 2021)

2.1 Abstract

Objective: To best meet our point-of-care research (POC-R) needs, we developed ProjectFlow, a configurable, clinical research workflow management application. In this article, we describe ProjectFlow and how it is used to manage study processes for the Diuretic Comparison Project (DCP) and the Research Precision Oncology Program (RePOP).

Materials and methods: The Veterans Health Administration (VHA) is the largest integrated health care system in the United States. ProjectFlow is a flexible web-based workflow management tool specifically created to facilitate conduct of our clinical research initiatives within the VHA. The application was developed using the Grails web framework and allows researchers to create custom workflows using Business Process Model and Notation.

Results: As of January 2021, ProjectFlow has facilitated management of study recruitment, enrollment, randomization, and drug orders for over 10 000 patients for the DCP clinical trial. It has also helped us evaluate over 3800 patients for recruitment and enroll over 370 of them into RePOP for use in data sharing partnerships and predictive analytics aimed at optimizing cancer treatment in the VHA.

Discussion: The POC-R study design embeds research processes within day-to-day clinical care and leverages longitudinal electronic health record (EHR) data for study recruitment, monitoring, and outcome reporting. Software that allows flexibility in study workflow creation and integrates with enterprise EHR systems is critical to the success of POC-R. Conclusions: We developed a flexible web-based informatics solution called ProjectFlow that supports custom research workflow configuration and

has ability to integrate data from existing VHA EHR systems.

2.2 Introduction

The Veterans Health Administration (VHA) is the largest integrated health care system in the United States, providing care for over 9 million Veterans at over 1255 facilities across the nation(V. H. A., n.d.). Over the past decade, the VA Office of Research and Development (VA ORD) has supported the implementation of point-of-care research (POC-R)(*Point of Care Research (POC-R)*, n.d.) with an emphasis on point-of-care clinical trials (POCCT), that is pragmatic clinical trials(Loudon et al., 2015) that compare approved treatment options when clinicians are in equipoise. The POC-R design aims to embed research workflows within day-to-day clinical operations with minimal interference to care, thereby allowing study endpoints, treatment deviations, and patient compliance data to be extracted from a patient's longitudinal electronic health record (EHR). POC-R also leverages EHR data for study recruitment. This clinically embedded design is advantageous in reducing research costs and promoting realization of a learning healthcare system by facilitating the translation of research evidence into clinical practice(Staa et al., 2012)(Vickers & Scardino, 2009). Workflow management software that can accommodate complex POC-R workflows while also integrating with enterprise EHR infrastructures is crucial to the success of point-of-care initiatives. Although numerous software options exist, many out-of-the-box tools lack flexibility(Nourani, Ayatollahi, & Dodaran, 2019), forcing researchers to either sacrifice integration with EHR systems or substantially modify their preferred study workflows. Within the VHA, software selection is further complicated

by requirements for compliance with robust data security rules as well as integration with its mature and rigid EHR infrastructures. Indeed, formal VHA approval for the installation and use of commercial or open source tools with protected health data may take years.

We developed a flexible web-based informatics solution called ProjectFlow that supports custom research workflow configuration and has ability to integrate data from existing VHA EHR systems. In this article, we describe how ProjectFlow is used to support a range of research study processes, from a focused patient enrollment process into a national data repository to the multiple complex processes of a multicenter POCCT.

2.2.1 Research Precision Oncology Program

Research Precision Oncology Program (RePOP) is the research component of the national clinical program named the Precision Oncology Program (POP). RePOP recruits and enrolls patients interested in sharing their de-identified electronic health data with the Precision Oncology Data Repository (PODR) for the purpose of innovating VHA cancer treatment(Do et al., 2019)(Elbers et al., 2020). This includes, but is not limited to, building genomic-based outcome prediction engines for clinical decision support. Cancer is a multifaceted disease, requiring an enormous amount of aggregate data to develop generalizable findings. To this extent PODR acquires patient genomic, clinical health, and medical image data (eg, radiological, pathological, and others). At present, RePOP utilizes ProjectFlow to support the contact and enrollment of eligible patients.

2.2.2 Diuretic Comparison Project

Diuretic Comparison Project (DCP) is a POCCT which evaluates the relative effectiveness of two widely prescribed thiazide-diuretics used in the treatment of hypertension. Specifically, the VHA national outpatient prescription database shows that over a million veterans were prescribed a thiazide-type diuretic each year from 2003 to 2008 with over 95% receiving hydrochlorothiazide (HCTZ) and less than 2.5% receiving chlorthalidone (CTD).(Ernst & Lund, 2010) However, recent evidence suggests CTD may not only be more effective for managing symptoms but also less expensive(Ernst & Lund, 2010). To better understand the relative effectiveness of these medications, the DCP trial compares cardiovascular outcomes in patients treated with HCTZ versus CTD. At present, DCP is actively enrolling patients and utilizing the ProjectFlow application to support trial recruitment, consent, randomization, and prescription order monitoring.

2.3 Materials and Methods

2.3.1 Key ProjectFlow functionalities

Customizable workflows and role-based access control

ProjectFlow is a flexible web-based workflow management tool created to facilitate the conduct of national POC-R within the VHA. ProjectFlow allows researchers to create custom workflows using the standards-based Business Process Model and Notation (BPMN, version 2.0). BPMN employs graphical elements as representations

of business processes(Ko, Lee, & Wah Lee, 2009)(White & Bock, 2011). Researchers may generate their study-specific workflows using the Eclipse IDE (Burnette, n.d.) and Activiti plugins(*Activiti*, 2022). ProjectFlow was developed using the Grails web framework. Previously known as "Groovy on Rails," Grails is an open source web application framework which uses Apache Groovy, a Java syntax compatible language(Judd, Nusairat, & Shingler, 2008). ProjectFlow also employs ″plug-and-play″ web-services to ensure smooth integration with EHR databases through synchronous (immediate response) or asynchronous (delayed re- sponse) communication. When creating POC-R workflows, ProjectFlow users first define "clinical elements." A clinical element is an object with specified properties that may move through a workflow. For example, a clinical element could be a physician, patient, medication, genomic sample or even a hospital location. Using BPMN 2.0 notation, users may design simple or complex workflows that a clinical element may move through (Figure 2.1A). The creation of study-specific workflows requires a general understanding of workflow notation as well as a moderate to strong understanding of database structures and relevant EHR data fields. Furthermore, a clear understanding of study-specific clinical processes is essential. Although initially a steep learning curve, we have found that staff members with sufficient training are able to contribute to complex workflow development within a few hours.

ProjectFlow employs Role-Based Access Control (RBAC) to manage how its users interact with clinical elements and workflows. RBAC allows control over which data a user (role) may access as well as the scope of functions (tasks) each role may perform. To keep patients' personally identifiable information secure, a study may restrict access to certain data elements depending on the role or expand access as the role

changes.

Integration with enterprise EHR systems

As noted previously, POC-R utilizes longitudinal EHR data for pa- tient recruitment and monitoring. ProjectFlow databases query VHA EHR systems regularly to obtain recent relevant data. The Vet- erans Health Information Systems and Technology Architecture (VistA), developed in 1999, serves as the backbone of the VHAâs EHR(Brown, 2003) until the transition to the Cerner EHR (fully operational by 2028) is complete. Clinicians access and enter patient data into VistA via the Computerized Patient Record System (CPRS) user in- terface. VistA data is transferred nightly to the VA Corporate Data Warehouse (CDW) where it resides in Microsoft SQL (MSSQL) databases for secondary use. Access to CDW data for research is managed by the VA Informatics and Computing Infrastructure (VINCI). To facilitate access to CDW systems the ProjectFlow appli- cation is hosted on virtual machines (VMs) within the VINCI envi- ronment. A generalized depiction of data flow for the ProjectFlow system is shown in Figure 2.1B.

Similar to our groupâs insulin POCCT pilot(D'Avolio et al., 2012)(L. D. Fiore et al., 2011) we have further leveraged VHA EHR functionalities through the use of modified CPRS/VistA interfaces. However, this time we have applied modifications on a national level as opposed to just a single VHA site. To provide a specific example, the DCP study utilizes CPRS ″View Alerts″ to facilitate provider consent to patient enrollment and randomization (Figure 2.2). As these alerts are embedded within the CPRS interface, the data entered within them becomes part of the patient 's longitudinal health record and the official system of record.

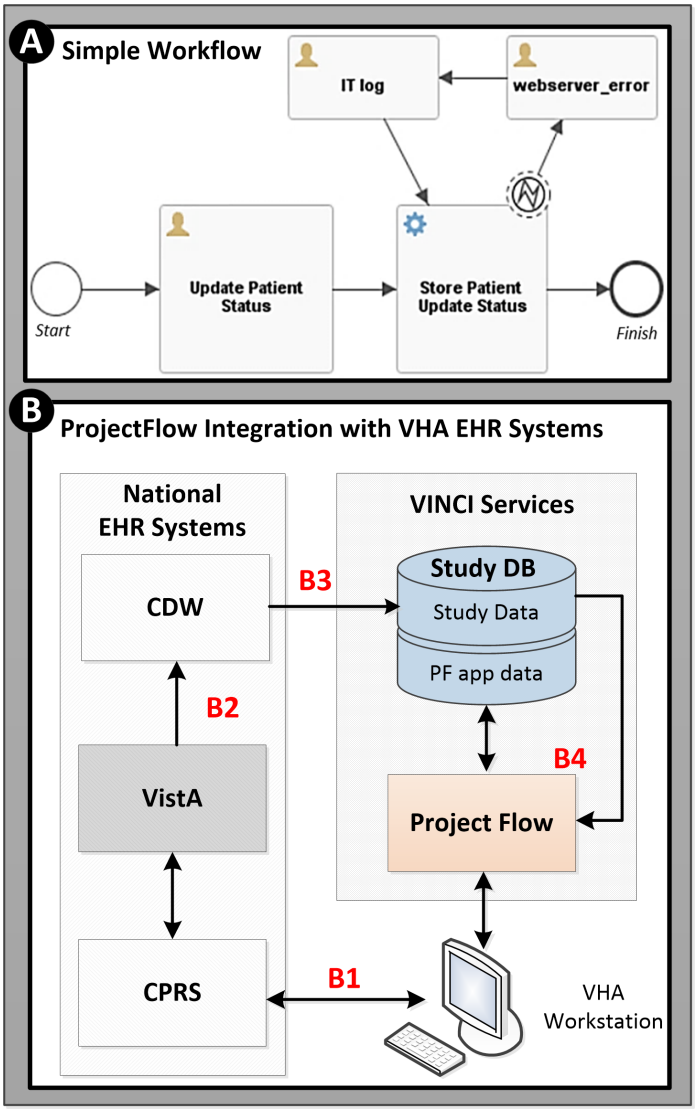


Figure 2.1: Workflow creation and integration with VHA EHR systems. (A) Example of simple BPMN 2.0 workflow: User defined clinical elements proceed through user designed study workflows. The figure depicts a workflow for updating a patient's status. More specifically, once the "Update Patient Status" task is completed, synchronous web-service communication transmits the updated information to the database. If an error occurs in transmission, this will be registered via the "IT logâ pathway. (B) Data flow utilized by the ProjectFlow web-based application: (B1) Clinicians utilize the computerized patient record system (CPRS) user interface to access and enter patient data into VistA. The DCP study utilizes "View Alerts" embedded within CPRS to facilitate trial recruitment, randomization, and prescription ordering. (B2) VistA data are transferred nightly to the VA Corporate Data Warehouse (CDW) where it resides for secondary operational and research use. (B3) Scheduled, nightly, extract transform load processes extract relevant EHR data from CDW into the study database (Study DB) which is utilized by (B4) ProjectFlow as needed for patient recruitment, randomization, prescription tracking and monitoring. Authorized study staff may access the ProjectFlow system and CPRS via their VHA workstation.

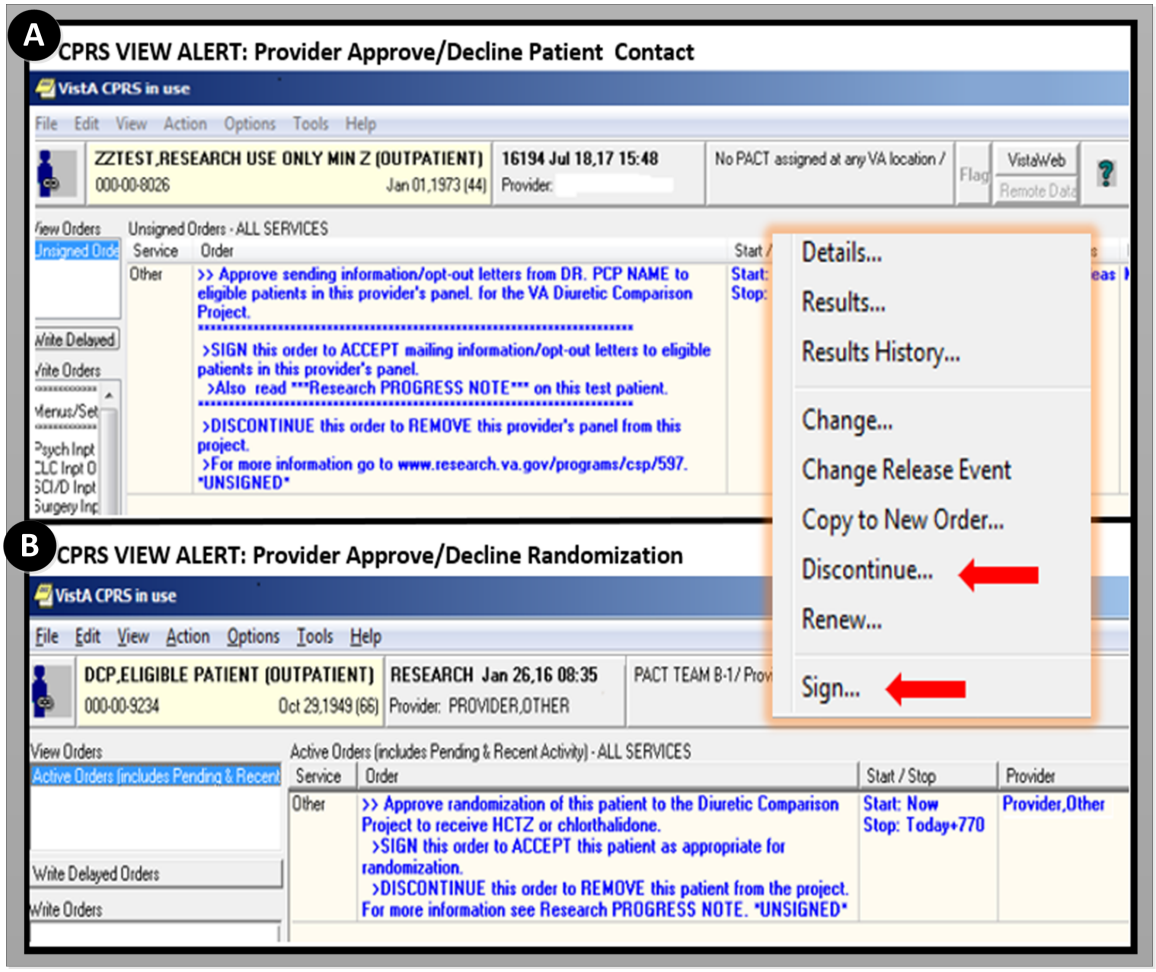


Figure 2.2: Integration with clinical interfaces, CPRS view alerts. (A) DCP utilizes CPRS "View Alert" screens to obtain provider consent to contact a patient (top panel) and (B) obtain provider consent to randomize consented patients (bottom panel). Providers may "Discontinue" or "Sign" these requests (red arrows). As the view alerts are embedded in CPRS the provider response also becomes part of the patient record. ProjectFlow queries and tracks these provider responses as they appear within the CDW.

ProjectFlow user dashboards and assigned task prioritization

Users interact with ProjectFlow by navigating within their dash- board (Figure 2.3). For example, Figure 3A shows what a user in the ″Nurse ″ role would see after selecting the clinical element ″Patient ″. ProjectFlow also helps users manage activities by prioritizing tasks associated with each of their assigned roles. For example, the ″Filter by Tasks ″ dropdown (Figure 2.3A2) lists the number of patient-related tasks that must be performed within each stage of a workflow. Prior to using ProjectFlow for study management, new staff undergo a 1.5-h training. Trainees must demonstrate a clear understanding of study workflows as well as how to navigate and manage their tasks using the application dashboards. For more details on dashboard functionalities, see the Figure 2.3 caption.

Pausing workflows and audit tracking

Clinical studies occasionally encounter unexpected situations that require additional discussion before proceeding. ProjectFlow allows clinical elements to be placed on ″HOLD ″ by flagging them within a workflow to allow time for issue resolution (Figure 2.3A4, red circle). When an element is placed on hold, the user may enter questions/comments or even assign follow-up to a different role. The role performing the follow-up may choose to provide the necessary information or may instead remove the element from the workflow entirely. Furthermore, to assure research integrity, clinical studies must maintain a record of actions for longitudinal auditing purposes. ProjectFlow facilitates record maintenance by logging critical information for each clinical element. Specifically, each element ’s entire path through a workflow is recorded, including but not limited to, the users performing the tasks, when each

Figure 2.3: ProjectFlow dashboard showing "Patient" clinical element views for the "Nurse" user/role. (A) Clinical elements appear at the top of the dashboard (red oval). In this example, a "Nurse" has already selected the element "Patient." (A1) The "All" tab lists tasks (complete, incomplete or on hold) assigned to any user/role. The "Assigned To Me" tab displays only tasks the Nurse role may execute. The Nurse has the ability to "release" the task after which it would appear in the "Patient Tasks" tab which lists all unassigned patient tasks. (A2) The "Filter by Tasks" dropdown shows which tasks require action by the Nurse as well as the number of patients for which that task must be performed. (A3) Data associated with a given patient or task may be queried using the search fields. (A4) In order to complete a task, the Nurse clicks the arrow ">" to open the "Complete Task" view. (B1) The Complete Task view displays relevant patient data. (B2) For this particular task the Nurse must decide whether or not to cancel the patients existing prescription order. (B3) After data entry, the task is completed by hitting the "Next" button. (B4) To view the workflow in its entirety and see which stage of the workflow a patient is currently in; the Nurse may select "View Workflow". (B5) The "History" panel lists completed workflow steps for the patient as well as which users completed them and when.

task is completed as well as all instances whereby an item was placed on hold or re-assigned to another user/role for completion.

2.3.2 Customized study workflows supported by ProjectFlow

Conceptual descriptions of the clinical study workflows supported by ProjectFlow for both RePOP and DCP are shown in Figure 2.4. As ProjectFlow accesses enterprise EHR data for provider and patient eligibility screening, both studies utilize it to automate and track study recruitment and enrollment.

RePOP workflows

RePOP recruits and enrolls patients interested in sharing their de- identified electronic health data with the Precision Oncology Data Repository (PODR) for the purpose of innovating (VA) cancer treatment(Do et al., 2019)(Elbers et al., 2020). To aid recruitment efforts, the ProjectFlow database has nightly extract transform load (ETL) processes that harvest the most recent available data for the RePOP cohort. The cohort consists of patients who, as part of their standard of care, have undergone or intend to undergo tumor-targeted genomic sequencing to further personalize their oncology treatment. Utilizing the ″ProviderRolodex″ workflow (Figure 2.4A, blue), providers are emailed study details and asked if their patients may be contacted to participate. If physician approval is granted, the ″PatientConsent″ workflow (Figure 2.4A, purple) is used to generate and manage postal mailings containing detailed study information, consent forms, and pre-paid return envelopes. The returned mail-

ings with signed consents are tracked by study staff in ProjectFlow.

DCP workflows

DCP is a POCCT which evaluates the relative effectiveness of two widely prescribed thiazide-diuretics used in the treatment of hypertension, that is HCTZ and CTD. DCP utilizes existing enterprise EHR data for provider and patient eligibility screening and recruitment (Figure 2.4B, blue, orange). First, EHR data housed in CDW is queried to identify qualifying providers and their associated living patients currently treated with HCTZ for hypertension. Providers' contact information is loaded into the trial DB and the ProjectFlow workflow "EmailProviders" is initiated (Figure 2.4B, blue). Providers are emailed study details and notified that they will receive a CPRS "View Alert", as exemplified in Figure 2.2A, to confirm or decline participation in the DCP trial; this process is managed and tracked via the "Physician Consent" workflow (Figure 2.4, orange). Providers may give permission for DCP staff to contact their qualifying patients about enrolling in the trial. Provider response to the View Alert, that is "sign" to approve, "discontinue" to decline, is subsequently captured by ProjectFlow via daily CDW extracts.

For consenting providers, a "Patient Search" algorithm pulls contact information for their eligible patients into the trial database and ProjectFlow application. The study Call Center then telephones eligible patients to inform them about the study and ask if they would like to participate. This patient contact and consent process is scripted and tracked within the ProjectFlow "Call Center" workflow (Figure 2.4B, green).

Consented patients are processed through the "RandomizePatient" workflow (Figure 2.4B, purple). Where relevant, physicians receive a view alert to approve or decline

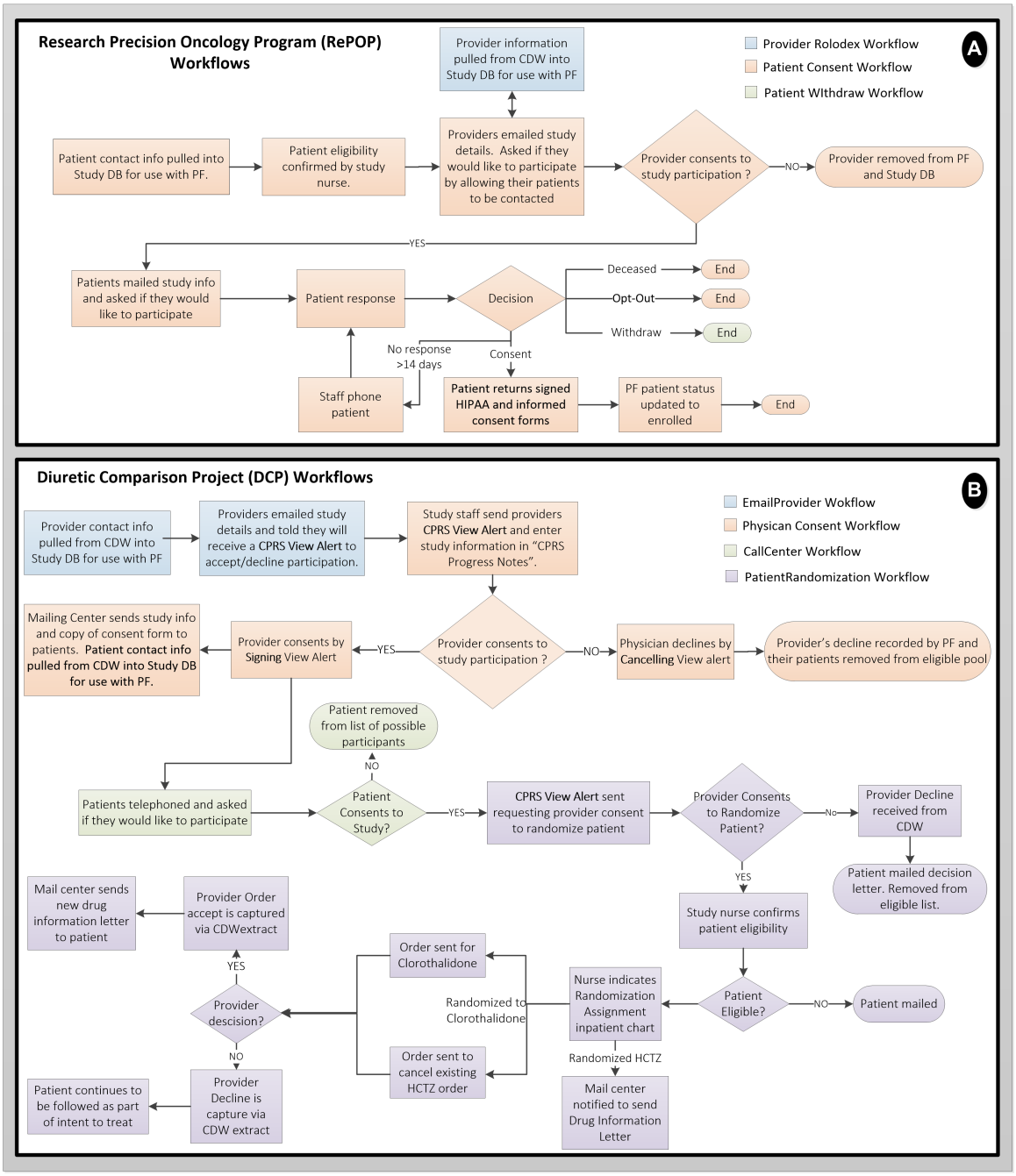


Figure 2.4: RePOP and DCP tasks managed and tracked by ProjectFlow. (A) RePOP workflows primarily support recruitment and enrollment (consenting) of patients. (B) DCP workflows not only support recruitment and enrollment but also study randomization and monitoring of prescription orders. Detailed descriptions of the workflows are provided in the main text.

patient randomization to CTHD (Figure 2.2B). Patients approved for randomization to CTHD are mailed detailed drug safety information. Those declined randomization are removed from the study eligibility list and sent a decision letter.

2.4 Results

2.4.1 Current status and application usage

The ProjectFlow application supports multiple study management functions and both RePOP and DCP have used the application to track recruitment and enrollment of providers and/or patients. It should be noted that patient enrollment numbers are most often related to the length of time a VHA site (hospital) has been participating in the study and not site-specific effort or engagement. To protect site confidentiality and assure continued enrollment safety, we do not specify VHA hospital locations (city, state) nor site-specific enrollment/randomization numbers.

Research Precision Oncology Program

Since 2017 over 11600 tasks have been executed in ProjectFlow by project managers, research nurses, and IT staff to track enrollment for RePOP at 41 VHA hospitals in 29 geographically distributed US states. As of January 2021, it has enrolled >370 patients and evaluated >3800 potentially eligible patients at 40 VHA hospitals across the United States. With the intent of enabling translational research to benefit our Veterans, de-identified, longitudinal clinical, targeted tumor sequencing, and medical imaging (CT and pathology slide) data from consented RePOP patients is available

to all researchers via the Veterans Precision Oncology Data Commons (VPODC)(Do et al., 2019)(Elbers et al., 2020).

Diuretic Comparison Project

Since 2017, DCP has utilized ProjectFlow to conduct and track over >230 000 tasks. To do this the application has been accessed >15000 times by call center staff, research nurses, project managers, and IT staff. It has also been used by the study call center to track phone calls made for recruitment as well as patient consent. As of January 2021, ProjectFlow has supported tracking of (and prescription ordering) for over 3500 providers and over 10 000 patients at 63 VHA hospitals across 39 geographically distributed US states.

2.5 Discussion

In contrast to explanatory trials, where the supporting infrastructure from patient selection to data collection is optimized to evaluate treatment efficacy, a pragmatic trial's infrastructure aspires to compare treatment effects with minimal interference in clinical practice; this often means relying on workflows that utilize the EHR. Indeed, merging research and clinical processes enables the use of longitudinal EHR data for study recruitment, enrollment, and monitoring. This can help reduce study costs and facilitate realization of a learning healthcare system through the direct translation of research evidence into clinical practice(Staa et al., 2012)(Vickers & Scardino, 2009). To successfully implement a pragmatic trial as a point of care clinical trial using the EHR requires flexibility and ease of integration that are not readily available in most

commercial and open-source trial management tools. Here, we have described our efforts with developing a software application that can support complex, customized workflows for POC-R task management while simultaneously integrating with VHA EHR infrastructures.

Initiated in 2016, the DCP study is one of VHA's largest POCCTs to date. Like other lengthy studies, DCP has evolved since its inception, including ongoing adjustments to the study protocol and overarching design. ProjectFlow's flexible workflow configuration has allowed us to adapt to these changing requirements and better manage more fluid tasks and successfully maintain continuity of study operations over-time. RePOP similarly has been able to iterate through protocol versions while the workflows in ProjectFlow were adapted accordingly. An example of this flexibility was a change from multi-level consent to one-level consent, which was accommodated by updating the variables in the consent steps to reflect the new consent form.

ProjectFlow's use of rule-based access control has also proven valuable in unexpected ways. Both RePOP and DCP studies have spanned several years and naturally undergone staff turn-over. Creating ″trainee″ roles with limited task scope and data access has helped new personnel learn our overarching POC-R study processes and workflows before moving into more formal, ″expert″ roles with greater scope and access. Similarly, the task prioritization feature of ProjectFlow's dashboards has not only helped experienced users manage their workload but increased new staff efficiency by serving as a reminder system that enhances learning. The application's ″Note" text-entry feature, within its "Hold″ functionality, has allowed staff to pause workflows and communicate on an individual case basis when questions arise. This has been particularly useful for DCP, which enrolls thousands of patients, as it can

reduce the need for separate email communication thereby helping maintain all necessary information within one auditing system.

In addition to customizable workflows, one of ProjectFlowâs most useful features is its ability to integrate seamlessly within the VHA EHR ecosystem. This minimizes disruption in clinical care while facilitating data management tasks. More specifically, for the DCP study, combined use of embedded CPRS View Alerts for data capture and CDW connections for data extraction greatly minimized the need for manual data entry by study staff and prescription monitoring via chart review. These features have also made ProjectFlow a useful recruitment tool for the VHA Integrating Pharmacogenetics in Clinical Care (I-PICC)(Brunette et al., 2020) trial aimed at assessing if genetic testing can effectively aid personalization of statin medications for cardiovascular disease treatment.

2.5.1 Limitations

ProjectFlow grants users significant flexibility in the types of workflows they may create. However, we have found that formal creation of BPMN workflows may require greater knowledge of database design and web-service protocols than the average trialist or clinical user may have. Thus, the overall maintenance and configuration of the application likely requires oversight by a software development team rather than being managed solely by trialists or clinicians. Additional work is required for reducing application complexity to allow management by the average user. Furthermore, as ProjectFlow's primary database is optimized to facilitate maintenance of the application itself, additional database structures designated for tracking and re-

porting study-specific outcomes on a project by project basis would be useful. More specifically, we have found that formal reporting may be complicated for clinical and reporting staff who are less familiar with the innerworkings of the application and workflows. Presently, to close this gap, database engineers create study-specific views and tables for reporting which may then be used by trialists for reporting.

2.6 Conclusion

POC-R aims to integrate research processes within day-to-day clinical operations and utilize patient EHR data for study recruitment, monitoring, and outcome reporting. The implementation of ProjectFlow to support both the DCP and RePOP studies afforded us significant flexibility to create customized study workflows for task management while simultaneously integrating with VHA enterprise EHR infrastructures. To date, ProjectFlow has facilitated management of study recruitment, enrollment, randomization, and drug orders for over 10 000 patients for the DCP clinical trial. It has also helped us evaluate over 3800 patients for recruitment and enroll over 370 of them into RePOP for use in data sharing partnerships and predictive analytics aimed at optimizing cancer treatment in the VHA. More recently, the application has also been used to support the VHA I-PICC trial.

2.7 Funding

Work for ProjectFlow development and implementation was supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research

and Development, Cooperative Studies Program. All work was conducted under approved IRB protocols for CSP2010 and CSP597. The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs nor the United States government

2.8 Author Contributions

Technical leadership for application development was provided by DE, NM, RA and RD. System development, quality assurance monitoring, development testing and database management was jointly performed by SLD, JF, QL, SS, RS, SD, SG, NM and TAF. All authors have contributed to aspects of this project.

2.9 Acknowledgements

We thank Ryan Cornia, Brian Ivie, Brad Adams, Lalindra DeSilva for their contributions to software creation and Pat Woods, Maura Flynn, Christal Sadatis, Amanda Guski, Cynthia Hau, Karen Pierce-Murray, and Corri Dedomenico for their help in development testing.

Chapter 3

The Veterans Affairs Precision Oncology Data Repository, a Clinical, Genomic, and Imaging Research Database

This Chapter is derived from Elbers & Fillmore et al. (Elbers et al., 2020)

3.1 The Bigger Picture

Accelerating the speed of innovation and discoveries in health care requires the liberation of real-world data from their silos. One of the greatest challenges in the application of artificial intelligence and machine learning to health care is the validation of new algorithms beyond where they were created. We present the Veterans Affairs Precision Oncology Data Repository (VA-PODR), a large-scale repository of de-identified data on patients diagnosed with cancer at the Department of Veterans Affairs (VA). VA-PODR includes longitudinal clinical, genomic, and imaging data originating from the VA's electronic health record system, the VA Central Cancer Registry, and other sources. VA-PODR enables researchers around the world to validate their algorithms and advance cancer research and health care in general. In addition, VA-PODR enhances Veterans' health care by facilitating development of algorithms that are well tuned to the Veteran population and ready for deployment inside the VA.

3.2 Summary

The Veterans Affairs Precision Oncology Data Repository (VA-PODR) is a large, nationwide repository of de-identified data on patients diagnosed with cancer at the Department of Veterans Affairs (VA). Data include longitudinal clinical data from the VAâs nationwide electronic health record system and the VA Central Cancer Registry, targeted tumor sequencing data, and medical imaging data including computed tomography (CT) scans and pathology slides. A subset of the repository is available

at the Genomic Data Commons (GDC) and The Cancer Imaging Archive (TCIA), and the full repository is available through the Veterans Precision Oncology Data Commons (VPODC). By releasing this de-identified dataset, we aim to advance Veterans' health care through enabling translational research on the Veteran population by a wide variety of researchers.

3.3 Introduction

The Department of Veterans Affairs (VA) operates the largest integrated health care system in the United States and was an early adopter of electronic health record (EHR) systems(Brown, 2003). As a result, the VA has large longitudinal databases containing health records dating back as far as the 1980s, with comprehensive coverage starting in 1999(*Corporate Data Warehouse (CDW)*, n.d.). More recently, the VA launched the Precision Oncology Program (POP), initially in the New England region under leadership of our group at the Boston Cooperative Studies Program (CSP) Informatics Center(L. D. Fiore et al., 2016)(L. Fiore et al., 2016), and subsequently as a national program led by the VA's National Oncology Program(Kelley, Duffy, Hintze, Williams, & Spector, 2017). Under POP, the VA has carried out targeted tumor sequencing on a national scale for Veterans diagnosed with cancer.

Although the VA has established successes in the care for cancer patients through evidence-based approaches, to accelerate this progress even further, it is critical to combine expertise from both inside and outside the VA. To that end, our group established the Research Precision Oncology Program (RePOP), which provides a mechanism for patients to consent to broad data sharing for research purposes. This, along

with additional work to assure regulatory compliance, enables the VA to contribute data into the cancer data ecosystem as recommended by the Cancer Moonshot Blue Ribbon Panel(Singer, Jacks, & Jaffee, 2016)("Blue Ribbon Panel Report", 2016). For example, consent obtained under RePOP has facilitated the VA's participation in the Cancer Moonshot's Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) network(*FACT SHEET: At Cancer Moonshot Summit, Vice President Biden Announces New Actions to Accelerate Progress Toward Ending Cancer As We Know It*, 2016).

Due to these efforts, we are now able to share VA data as a national resource to investigators inside and outside the VA. Specifically, in this paper, we introduce the VA Precision Oncology Data Repository (VA-PODR) and describe its availability outside the VA. VA-PODR consolidates de-identified VA clinical, genomic, and imaging data needed for research in precision oncology in a large, nationwide repository. The data consist of longitudinal clinical data from the VA's integrated EHR system and the VA Central Cancer Registry, targeted tumor sequencing data, and medical imaging data including computed tomography (CT) scans and pathology slides.

In previous work, we described the VPODC, a data-sharing and computational platform where VA-PODR is available to researchers outside the VA(Do et al., 2019). Here, in contrast, we describe the VA-PODR data repository itself. The difference between the VPODC and VA-PODR is that the VPODC is a specific platform through which the VA-PODR dataset is shared and collaborative analysis can take place. However, the VPODC is not the only place where VA-PODR is available. Parts of VA-PODR are also shared in the Genomic Data Commons (GDC)(*Genomic Data Commons*, n.d.) and The Cancer Imaging Archive (TCIA)(Clark et al., 2013), as well

as internally at the VA. And, in the future, it is our intention that VA-PODR will be made available elsewhere as well.

VA-PODR represents a unique resource, for several reasons. First, to our knowledge, VA-PODR is the first large-scale repository of VA health data that has ever been made available to researchers outside the VA. Second, VA-PODR is one of only a handful of large-scale databases with real-world EHR data to be made available to the research community outside the institutions where the data originated, similar to the widely used MIMIC Critical Care Database(Johnson et al., 2016), but in a different medical domain. Third, VA-PODR combines clinical, genomic, and imaging data, enabling multimodal analysis that is not possible with only one source, similar in some respects to The Cancer Genome Atlas (TCGA)(Weinstein et al., 2013), but with substantially richer clinical data.

In addition to VA-PODR's use for observational research in precision oncology, we envision VA-PODR facilitating big data research in health care in general, such as for the development and validation of analytical models and other tools using large EHR, genomic, and imaging data, and also being used for educational purposes. Thus, our contribution will benefit both Veterans and the broader community by providing accessible data to accelerate improvement in health care.

3.4 Results and Discussion

3.4.1 Methods

Repository Development

VA-PODR consists of (1) de-identified longitudinal clinical data from the VAâs integrated EHR system and the VA Central Cancer Registry, (2) targeted tumor sequencing data from POP, and (3) medical imaging data including CT scans and pathology slides. Data from multiple sources is pulled to a central location, and records are matched using internal patient identifiers (Figure 3.1).

Clinical data include 10 domains, 38 tables, and 647 columns, and approximately 1 billion rows of data from the VA's EHR system, with detailed information on demographics, survival, laboratory test results, orders, medications, surgeries, and visits, including associated ICD and procedure codes, and more.

This information is available in the VA Corporate Data Warehouse's native data model as well as in the VA's implementation of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Our intention is for VA-PODR to include all information from the EHR that is relevant for research. In addition, extensive curated data on cancer diagnosis, treatment, and outcomes is included from the VA Central Cancer Registry (VACCR)(Zullig et al., 2012), which collates data on cancer cases that have been annotated by VA cancer registrars throughout the nation.

Genomic data include data generated under both the VA New England Healthcare System Precision Oncology Program(L. Fiore et al., 2016)(L. D. Fiore et al., 2016)

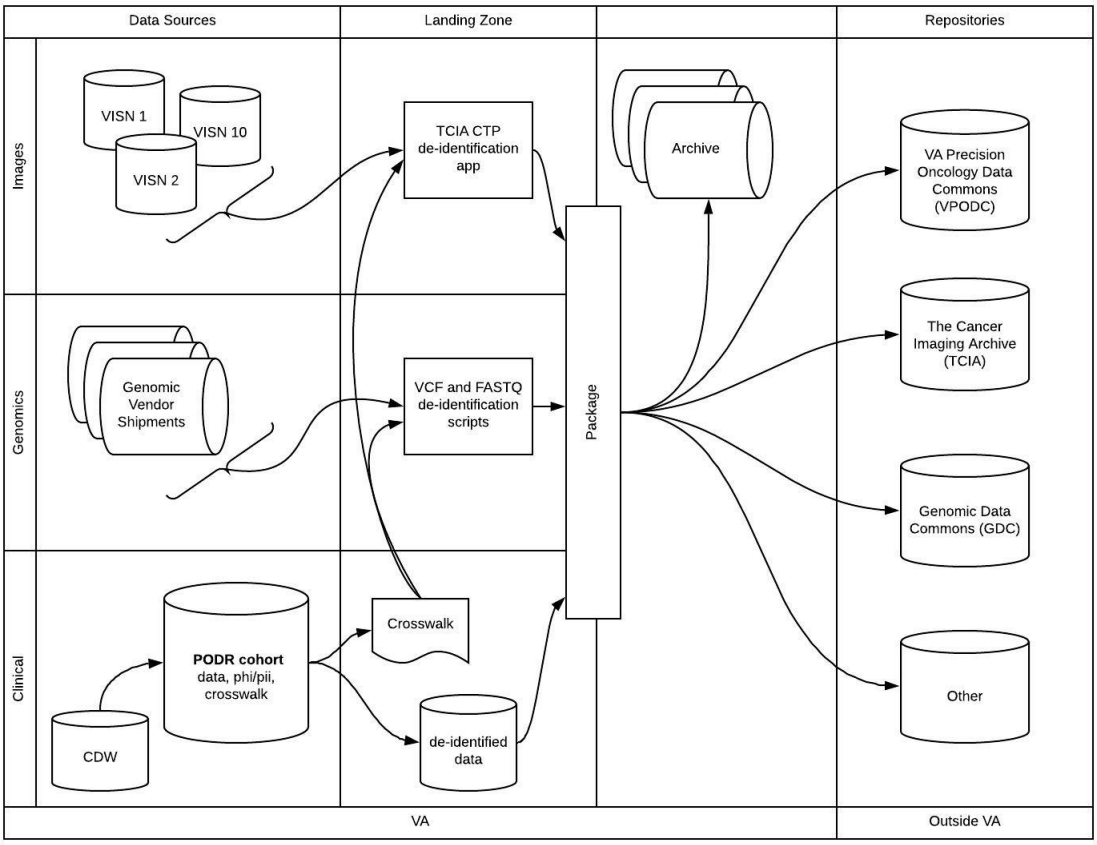


Figure 3.1: Overview of the VA-PODR Dataflow. Data are pulled from several sources within the VA, aligned and de-identified in the landing zone, and subsequently submitted to collaborating repositories.

and the VA National Precision Oncology Program(Kelley et al., 2017). These data consist of targeted tumor sequencing data, including both raw sequencing files and somatic variant calls.

Medical imaging data are extracted from two distinct sources. CT scans taken at or near diagnosis are pulled from various VA's medical centers' Picture Archiving and Communication Systems. In addition, images of pathology slides from sequenced tumor biopsies are included; these were produced by the vendors carrying out the targeted tumor sequencing described above, or are histopathology slides digitized at local pathology departments.

De-identification

General Strategy Before release, all data are de-identified in accordance with both the United States Health Insurance Portability and Accountability Act (HIPAA)(Office for Civil Rights, 2012) and internal VA requirements, including VHA Handbook 1200.12 guidelines(*ORD VHA Directive, Handbooks, and Program Guides – 1200 series*, 2022). Specifically, all data elements covered under HIPAA as identifiers are removed, as well as other possible sensitive information, identified as such by subject matter experts (SMEs); see details below. After de-identification using the procedures below, all data are reviewed and approved by a VA Privacy Officer before release.

In all data types, dates are obfuscated by calculating days to an arbitrary patient-specific anchor date. This procedure preserves the ordering of dates within each patient's timeline. No patient data are included for which the timestamp indicates that it was collected at the time a patient is 90 years of age or older.

Clinical Data

As briefly mentioned above, the clinical data are de-identified through a manual review for identifiers under HIPAA, sensitive information, and/or other PHI/PII by SMEs. Specifically, each column of data is assessed for inclusion or exclusion based on whether it primarily contains this type of information, which can occur in categorized fields or free text, or not. Columns that primarily contain sensitive information are excluded. In addition, for columns that are included, each distinct data value is evaluated for sensitivity by an SME and either excluded or added to a white list. Examples of sensitive data are medication discontinued dates, serial codes, and references to physicians and locations. As new data are added, they are checked against the white list and flagged for manual review if not present (Figure 3.2). Thus, SMEs are involved in two steps: first in assessing each new column and deciding if the column in its entirety contains sensitive and/or HIPAA information and second in reviewing new and distinct varchar values and deciding if this specific value should be white- or blacklisted while the column itself is part of PODR.

Genomic Data

Genomic data are de-identified by removing any HIPAA identifiers occurring in metadata associated with each record and certifying this de-identification methodology by an SME. Genomic data of tumors are not considered inherently identifiable under US regulations(Dankar, Ptitsyn, & Dankar, 2018).

Imaging Data

For imaging data, we ensure that both headers and image content do not include

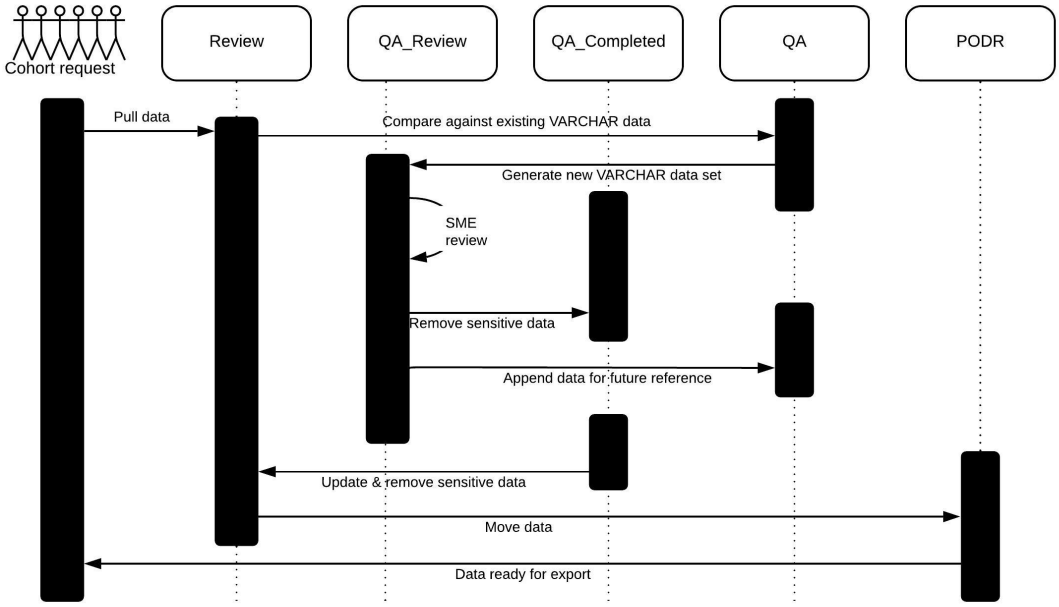


Figure 3.2: Review Process to Exclude Sensitive and Identifiable Data. Once a cohort is requested, new data are pulled and unique values are compared with data values previously evaluated. New data values are evaluated by SMEs and either white- or blacklisted. The new dataset is filtered by the white-listed dataset, de-identified, and shared.

sensitive or identifying information. Sensitive data in Digital Imaging and Communications in Medicine (DICOM) headers are removed through a process similar to that used for clinical data. To de-identify image content, we use the Medical Imaging Resource Community's Clinical Trials Processor software(Langer et al., 2015).

Technical Validation

Data have been validated three ways. Best practices have been used in developing code to pull and collate data, relying on standard internal VA identifiers to join data across domains where possible, and subsequently translating these to public identifiers. In addition, all data domains and values have been reviewed by a group of SMEs with expertise in medicine, imaging, and data analysis. Finally, all data have been reviewed and certified as de-identified by a VA Privacy Officer before release.

Risk of Re-identification

In sharing VA-PODR data, de-identification standards have been adopted, extended, and thoroughly vetted to protect VA patients. In order to further mitigate residual re-identification risk in the de-identified datasets, we have also implemented policy controls. Specifically, the larger set of de-identified data is available only after proof of institutional privacy policies, vetting of data transfer rules, and signing of a Data Use Agreement (DUA), in which the user agrees not to attempt re-identification. These measures are further detailed in the next section.

Regulatory Considerations

In accordance with VA regulations, VA data can be housed in a database, a single repository, or a data repository. A database is explicitly created in the course of conducting research, while a single repository can be used to archive/compile data from multiple protocols and/or investigators working on similar topics. VA-PODR is a data repository, meaning that it is a living document that details the sources and contents of archived research data and also describes all secondary use of data including where the data go, who uses them, for what purposes, over time. VA-PODR currently describes two cohorts, A and B, with their regulations detailed below. We anticipate other cohorts will be added to VA-PODR in the future, whether prospective or retrospective, observational or interventional, of which APOLLO is an example.

Cohort A Consent and HIPAA Authorization

For patients in cohort A, consent and HIPAA authorization was obtained to share data with partners inside and outside the VA. Inside the VA, data on patients in cohort A can be shared either identified or de-identified while coded or anonymized, depending on the investigator's protocol and the details of the DUA signed with PODR. Outside the VA, data on patients in cohort A can only be shared after they are de-identified and coded or anonymized and a DUA is signed with PODR. This DUA may allow data on patients in cohort A to be reshared. The signed DUA will state that re-identification is prohibited.

Cohort B Decedent and HIPAA Waived

Patients in cohort B need to be recorded as decedent in their medical record; iden-

tification of these patients is performed under an HIPAA waiver. Decedent data are not considered data on human subjects. Similar to cohort A, inside the VA, data on patients in cohort B can be shared either identified or de-identified while coded or anonymized, depending on the investigator's protocol and the details of the DUA signed with PODR. Outside the VA, data on patients in cohort B can only be shared after they are de-identified and coded or anonymized and under a DUA signed between PODR and a trusted partner with a standard operating procedure or institutional review board-approved protocol. Different from cohort A, these data cannot be reshared and must remain in the trusted partner's repository. The signed DUA will state that re-identification is prohibited.

Unanticipated Events

If an unanticipated event occurs with PODR data, the procedures in the most recent version of the PODR protocol and VA handbook 1200.05 are followed. The event is reported to the privacy officer, principal investigator, quality assurance manager, institutional review board, and RD committee at the VA Boston Healthcare System. In addition, the event is entered into a national privacy database by the privacy officer, and a determination is made on the need for corrective action.

3.4.2 Results

Patient Characteristics

In its current release, VA-PODR includes data on 113,154 Veterans diagnosed with cancer at the VA, including 1,115 patients who were enrolled in POP. Details on

Characteristic	VA-PODR (N = 113,154), n (%)
Age	
<50 years	391 (0.4)
50-59 years	7,215 (6.4)
60-69 years	38,170 (33.7)
70-79 years	37,151 (32.8)
≥ 80 years	30,227 (26.7)
Gender	
Male	111,811 (98.8)
Female	1,343 (1.2)
Race	
African American	20,531 (18.1)
American Indian	512 (0.5)
Asian	247 (0.2)
White	77,295 (68.3)
Native Hawaiian/Pacific Islander	668 (0.6)
Unknown	13,901 (12.3)
Ethnicity	
Hispanic or Latino	3,624 (3.2)
Not Hispanic or Latino	99,862 (88.3)
Unknown	9,668 (8.5)

Table 3.1: Demographic Characteristics of the VA-PODR Patient Population. (Patients can report multiple races.)

demographics, date of diagnosis, and cancer type are shown in Tables 3.1, 3.2, and 3.3. Patients in VA-PODR are predominantly older (26.7% ≥ 80 years, 32.8% 70-79 years, 33.7% 60-69 years, 6.4% 50-59 years, and only 0.4% <50 years) and male (98.8% male, 1.2% female). The cohort includes 18.1% African American and 68.3% white patients. Cancer types include prostate (58,323 patients), lung (56,836), bladder (3,640), skin (2,168), colon (2,043), and kidney (1,284) cancer, among others. Year of diagnosis ranges from 2005 to 2019, with 2.6% of diagnoses in 2004 or earlier.

Cancer Type	VA-PODR (N = 113,154), n (%)
Prostate	58,323 (51.5)
Lung	56,836 (50.2)
Bladder	3,640 (3.2)
Skin	2,168 (1.9)
Colon	2,043 (1.8)
Kidney	1,284 (1.1)
Other	11,963 (10.5)

Table 3.2: Distribution of Cancer Types in the VA-PODR Patient Population as Reported by the VACCR.(Patients can report multiple races.)

Year of Diagnosis	VA-PODR (N = 113,154), n (%)
$\leq$ 2004	2,986 (2.6)
2005	7,158 (6.3)
2006	6,658 (5.9)
2007	6,918 (6.1)
2008	6,004 (5.3)
2009	6,752 (6.0)
2010	12,517 (11.1)
2011	11,710 (10.3)
2012	10,742 (9.5)
2013	9,784 (8.6)
2014	9,203 (8.1)
2015	8,316 (7.3)
2016	6,698 (5.9)
2017	5,080 (4.5)
2018	2,336 (2.1)
2019	106 (0.1)
Unavailable	270 (0.2)

Table 3.3: Year of Diagnosis of Cancer in the VA-PODR Patient Population as Reported by the VACCR

Efforts are underway to expand VA-PODR to include all lung cancer patients within the VA since 1999 (approximately 150,000 cases) and add at least 100,000 prostate cases. Ultimately, we intend all known cancer cases at the VA to be included in VA-PODR. We are also conducting iterative updates of images and genomic data as they become available, as well as expanding clinical data domains to, for example, include the inpatient and radiology domains. We are open to input from the research community on determining expansion priorities.

Classes of Data

Clinical data are stored in a relational database for the following domains: outpatient visits, inpatient medications, outpatient medications, all laboratory test results, all orders, surgery, and patient demographics. Clinical data also include cancer registry data and a derived table containing all ICD codes and timestamps associated with each patient across several domains (Table 3.4).

Genomic data are stored in individual files, linked to clinical data by filename identifiers. Genomic data include three levels of detail: raw sequencing data in FASTQ format, somatic mutation data stored in VCF format, and information on actionable mutations based on curation by molecular diagnostic vendors. Two vendors were used: Personalis, which used the ACE Cancer Plus panel, and PGDx, which used the Cancer Select 203 panel. Information on which vendor was used for each patient sample is available in the metadata.

Imaging data are stored as DICOM stacks in Orthanc(Jodogne, 2018), an open source PAC server, and are linked to the clinical data by patient identifiers. A set of the metadata extracted from the DICOM tags is available in table format for direct

Domain Table Name	Description
BCMAMedicationLog	Inpatient medications
CPRSOrder	All orders for drugs, labs, etc.
OutpatVDiagnosis, OutpatVisit, OutpatVPatientEd, OutpatVProcedure, OutpatVProcedureDiagnosis, OutpatVSkinTest, OutpatVSkinTestDiagnosis	Outpatient visit information
PatientLabChem	Laboratory test information
PatientMeansTest	Patient income information
Patients	Basic demographic and vital status information
RxOutpat	Pharmacy outpatient
SurgeryPRE, SurgeryINTRA, SurgeryPOST, SurgeryProcedureDiagnosisCode	Surgery data
ICDCode	All ICD codes with timestamp
OncologyPrimary	VA Central Cancer Registry Data
OMOP	The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

Table 3.4: Data Domains Available in VA-PODR

querying.

Clinical Data Domains

VA-PODR contains a broad range of longitudinal clinical data both from the VA's EHR system and from the VACCR. EHR data in VA-PODR include information on patient demographics, comorbidities, procedures, medications, laboratory test results, medical orders, survival, as well as detailed administrative information arising from inpatient and outpatient visits to the VA. In addition, VACCR data in VA-PODR include extensive manually annotated information on cancer cases and outcomes.

Repository Access and Locations

VA-PODR at present consists of two different cohorts with different access policies. Cohort A includes patients who have consented to broad data sharing with parties external to the VA under the RePOP protocol. Cohort B includes patients who are deceased and are not considered human subject research, but by internal policy is subject to additional restrictions compared with cohort A. In the future, additional cohorts may be added.

A subset of VA-PODR data from patients in cohort A is available at the Genomic Data Commons(*Genomic Data Commons*, n.d.)(Grossman et al., 2016) and TCIA(Clark et al., 2013)(*TCIA Collections*, n.d.) and our intention is that relevant data elements from all of cohort A will be available at these locations in the future. All VA-PODR data for cohort A are available at the Veterans Precision Oncology Data Commons (VPODC)(*The Veterans Precision Oncology Data Commons*, n.d.) to approved users who provide proof of institutional privacy policies and data transfer rules to the data steward. Cohort B is also accessible at the VPODC through the data steward and with the approval of the data owner (VA). Like the GDC, the VPODC is a platform for data sharing and analysis in the Gen3 commons framework developed by the University of Chicago(Grossman et al., 2016).

To request access to VA-PODR via VPODC, users should send an access request email to the contact listed at https://vpodc.org. The initial email should include a description of the research topic, the cohort to which access is requested, and institutional details. Access requests will be reviewed and prioritized by an allocations committee. Before data access is granted, all users will be required to complete training on privacy and information security. In addition, users will be required to

sign the applicable and most recent version of the DUA.

FAIR Principles

VA-PODR aligns with the FAIR principles(Wilkinson et al., 2016) that data should be findable, accessible, interoperable, and re-usable. A full data dictionary is included in VA-PODR and available at the VPODC. In the GDC and TCIA, the data follow the data structures mandated and documented by those frameworks, creating a findable and rich metadata structure with unique and persistent identifiers. The VPODC, GDC, and TCIA follow interoperable standards for authentication, access, and retrieval. The subset of VA-PODR data that follows the OMOP data structure allows for interoperability with other health care data repositories, and the subset of data available at GDC and TCIA follows those systems' standards. Since data will be housed long-term at these sites, it will remain findable and re-usable. Thus, we believe that the VA-PODR satisfies the conditions of being FAIR.

Use Cases

In addition to VA-PODR's primary use to enable research in precision oncology, the dataset has substantial use for several other purposes of broad interest, including (1) methods research in analysis using EHR data, (2) educational and training purposes, and (3) development of clinical informatics tools. VA-PODR is particularly valuable for these additional purposes, because there are currently few large EHR datasets readily available to researchers or students who are not affiliated with an institution that has an EHR data warehouse.

To date, VA-PODR data have been used in the Department of Commerce's The

Opportunity Project (TOP) Health Artificial Intelligence initiative(*Deep Dive: How a Health Tech Sprint Pioneered an AI Ecosystem*, 2019)(*TOP Health Sprint*, n.d.) In addition, VA-PODR data have been used to carry out external validation and calibration of a prognostic model for mortality among patients with non-small cell lung cancer(Cheng et al., 2019)(N. Fillmore et al., 2019). In addition, research projects on lung and prostate cancer using VA-PODR data and the VPODC platform are underway.

3.4.3 Conclusion

We have described VA-PODR, a large, nationwide repository of de-identified data on patients diagnosed with cancer at the VA. This repository contains longitudinal clinical data from the VA's nationwide EHR system and the VACCR, targeted tumor sequencing data, and medical imaging. A subset of the repository is available at the GDC and TCIA, and the full repository is available through the VPODC. By making these data available, VA-PODR enables multiple uses of benefit to both Veterans and the broader community.

3.5 Experimental Procedures

3.5.1 Resource Availability

Data and Code Availability

VA-PODR is available at https://vpodc.org, and subsets of the data are available at https://portal.gdc.cancer.gov/projects/VAREPOP-APOLLO and https://wiki.cancerimagingarchive.net/display/Public/APOLLO-1-VA.The published article reports on all data generated to date (July 25, 2020), and it is anticipated that the data will be expanded. There are restrictions to the availability of data due to privacy considerations, as described above.

Acknowledgments

This work was supported by the VA Office of Research and Development, Cooperative Studies Program (CSP). The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Author Contributions

D.C.E. and N.R.F. cowrote the paper. D.C.E., N.R.F., F.M., and D.C.C. designed the VA-PODR adjudication and de-identification methodology. D.C.E., F.-C.S., and S.S.G. designed VA-PODR's architecture and executed its development. A.P., C.M., and R.L.G. designed and implemented VPODC. S.J.A. provided regulatory guidance.

M.T.B., N.V.D., and R.L.G. supervised the work. All authors contributed to VA-PODR and critically reviewed the paper.

Chapter 4

An application to support COVID-19 occupational health and patient tracking at a Veterans Affairs medical center.

This Chapter is derived from Fillmore & Elbers et al. (N. R. Fillmore et al., 2020)

4.1 Abstract

Objective: Reducing risk of coronavirus disease 2019 (COVID-19) infection among healthcare personnel requires a robust occupational health response involving multiple disciplines. We describe a flexible informatics solution to enable such coordination, and we make it available as open-source software. Materials and Methods: We developed a stand-alone application that integrates data from several sources, including electronic health record data and data captured outside the electronic health record. Results: The application facilitates workflows from different hospital departments, including Occupational Health and Infection Control, and has been used extensively. As of June 2020, 4629 employees and 7768 patients and have been added for tracking by the application, and the application has been accessed over 46 000 times. Discussion: Data captured by the application provides both a historical and real-time view into the operational impact of COVID-19 within the hospital, enabling aggregate and patient-level reporting to support identification of new cases, contact tracing, outbreak investigations, and employee workforce management. Conclusions: We have developed an open-source application that facilitates communication and workflow across multiple disciplines to manage hospital employees impacted by the COVID-19 pandemic.

4.2 Introduction

The incidence of coronavirus disease 2019 (COVID-19) in healthcare personnel has been reported to vary between 3.8% and 38.9%(Chou et al., 2020)(CDC COVID-19

Response Team, 2020). Multiple strategies have been proposed to manage and protect these critical individuals during the pandemic(Ehrlich, McKenney, & Elkbuli, 2020)(Adams & Walls, 2020)(Shanafelt, Ripp, & Trockel, 2020)(Semple & Cherrie, 2020)(Nagesh & Chakraborty, 2020)(Godderis, Boone, & Bakusic, 2020)(Gan, Lim, & Koh, 2020). Risk reduction strategies involve using policies, processes, and technologies to address screening, testing, monitoring, and quarantining. Characterizing the workforce by occupation, location, symptoms, exposures, and test results is critical to formulate and implement risk reduction strategies, and these data vary over time(Adams & Walls, 2020)(Adalja, Toner, & Inglesby, 2020)(Baker, Peckham, & Seixas, 2020). The fragmentation of information challenges occupational health staff, infection prevention nurses, and clinical providers to effectively understand the pandemic's impact on the workforce of an institution in real time. There is a growing number of technology platforms and applications to augment the workflow of screening the patient population, tracking infections, monitoring supplies, and self-triage to support clinical operations(Dong, Du, & Gardner, 2020)(Judson et al., 2020)(Vaishya, Haleem, Vaish, & Javaid, 2020)(Berry, Soucy, Tuite, & Fisman, 2020). There is also an increasing number of applications that focus on categories such as âdiagnosis, prevention, treatment, adherence, lifestyle, and patient engagementâ(Patel et al., 2020)(Hollander & Carr, 2020)(Reeves et al., 2020)(Golinelli et al., 2020). Although these technologies tackle many important public health concerns, they have not provided adequate support for a comprehensive healthcare system's occupational health response to the pandemic.

From March through May of 2020, Boston was among the most significantly affected metropolitan areas in the United States with new COVID-19 cases. Several

organizations such as The New York Times and Johns Hopkins University report data on COVID-19 cases and related deaths, facilitating understanding of the epidemic at a national and state level. Dashboards such as the COVID-19 Watcher have been created to quickly ascertain information such as the number of cases in Boston and how the city compares to other cities as regards to prevalence of COVID-19(Wissel et al., 2020). However, it is extremely difficult to quickly determine how many patients and employees are diagnosed with or potentially to COVID-19 infections on a daily basis at an institutional level. Applications have been developed that use information and configurable tools from the electronic health record (EHR) to track persons under investigation and infected patients in hospitals(Reeves et al., 2020), but questions related to potential exposures and testing of employees are often difficult to answer quickly because complete employee health records are not routinely found in the hospital's EHR.

4.3 Objective

We describe our COVID-19 Data Management Platform (COVIDDMP), a flexible informatics solution to facilitate communication and workflow across multiple disciplines in response to the rising occupational exposure to COVID-19. We discuss how our application can quickly adapt to the evolving policies and procedures for protecting healthcare workers as the pandemic progresses. The application is available as open-source software for use by other healthcare systems (https://github.com/bostoninformatics/COVIDDMP). The open-source approach allows us to create a flexible framework that can be adapted to integrate data from

different sources and interoperable with a range of EHR platforms.

4.4 Materials and Methods

The Veterans Affairs Boston Healthcare System (VABHS) comprises 3 campuses and 5 Community-Based Outpatient Clinics, providing a comprehensive range of services including primary and specialty care, surgical and emergency services, and short-term and long-term community living centers. Like other healthcare centers throughout the Department of Veterans Affairs (VA), VABHS uses the Veterans Health Information Systems and Technology Architecture (VistA) as the EHR(Brown, 2003).Real-time access to VistA data by third-party applications is tightly controlled, but VistA data from all VA medical centers are collated in a large data warehouse, the Corporate Data Warehouse (CDW), and updated nightly(Fihn et al., 2014). Given the rapid development cycle and the need for real-time data, instead of seeking changes to VistA, we chose to develop a stand-alone, open-source platform that integrates data captured outside of the EHR and data exported from the EHR. This added the capability of a 3 times per day VistA extract, utilizing FileMan("VA FileMan Technical Manual", 2021), the database management system of VistA, to extract near real-time reports, and user-entered data, to the nightly data extract from CDW. Clinical stakeholders determined a 3 times per day frequency of FileMan reports as being adequate for their workflow. Our approach combines a Python-based engine and graphical user interface in the form of a dashboard(Grinberg, 2018). The application is enhanced with access controls and security and integrates with an underlying pipeline that accomplishes data source harmonization and integration using R (version 3.6.3, R Foundation for

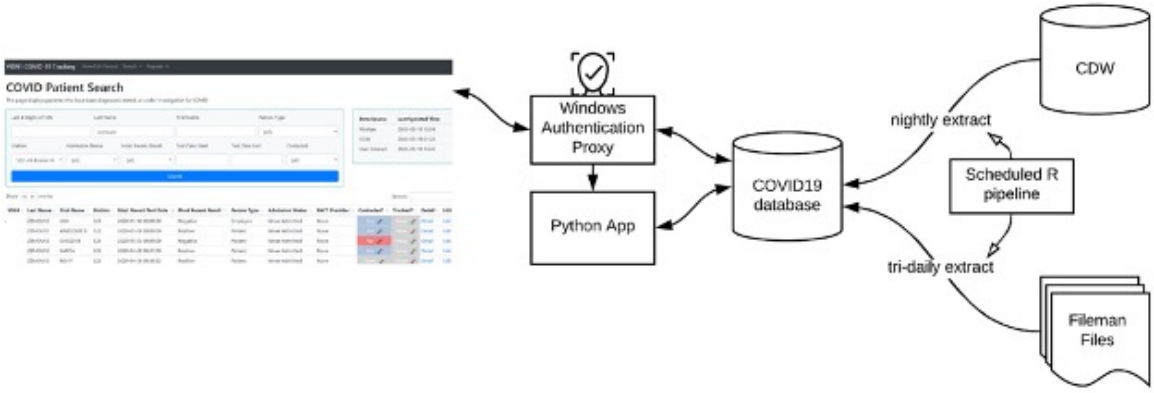


Figure 4.1: System architecture. A scheduled R pipeline extracts data from the Corporate Data Warehouse (CDW) nightly and from a FileMan export thrice daily, and loads this data in the application's COVID19 database. The front end, including a Windows Authentication proxy and the Python-based web application, read from the application database and serve results to the user.

Statistical Computing, Vienna, Austria) and SQL pipelines (Figure 4.1).

In order to develop and deploy COVIDDMP while policies and procedures are evolving, a rapid and dynamic development structure was put in place. This involved early morning 30-minute scrums with the development team and late afternoon 30-minute check-ins with our stakeholders to update on progress and clarify outstanding requirements. The development team consisted of data scientists, front-end and database engineers, report and quality assurance specialists, and system administrators led by technical architects and project managers. The stakeholders were clinicians, nurses and technical support staff from VABHS, primarily from occupational health, but other teams such as primary care and infectious disease were involved as well. To track requirements, implement a software development life cycle, and facilitate version control, the VA's GitHub platform was utilized. This allowed technical architects and project managers to create issues describing and prioritizing the requirements and bugs, while developers check in their code and branch out to

build features. When features passed quality assurance the code was merged back in after careful review in a pull request by one of the technical architects. During the initial phases of deployment, daily deployment occurred, which stabilized over time to a weekly deployment cycle.

4.5 Results

4.5.1 Application Components

The resulting COVIDDMP consists of 4 main parts: the data extraction pipeline, the database, the GUI application, and the integrated authentication. The data extraction pipeline has been written in R and SQL. It runs overnight to extract new CDW data and every 15 minutes between 7 am and 7 pm to verify if new FileMan data are available to be integrated into the COVID-19 database. The data extracted from CDW and the FileMan report include new tests, test results, and persons under investigation. In addition to this, the pipeline pulls data on the patient or employee including contact information, hospitalization status, ward location, and the Patient Aligned Care Team. The Microsoft SQL Server (version 11.0.7001, Microsoft Corporation) database maintains this extract, assigns unique person identifiers to maintain data integrity, and stores a history of all data entered manually by the user. Users are able to add new persons to the database and update or correct information on existing persons. The COVIDDMP assumes that data entered or corrected by the user takes precedence. If more recent data are available from the extract, this is shown separately. This allows users to update the data in real time and not have to

wait until the next data pull. This increases efficiency and improves clinical decision making for the healthcare personnel accessing the database. To support rapid development and changing requirements, the COVIDDMPâs database table structure follows a relatively flat, non-normalized model, relying on view structures to provide the most recent data.

The Python-based GUI application contains Search, Add/Edit, and Reporting tabs. The Search tab allows users to search for their person(s) of interest based on dropdown variables such as type (eg, patient or employee) or ward location and personally identifiable information. Frequently used settings are prepopulated and available as options under the Search tab. Additionally, the Search tab allows the user to track when persons were contacted, drill down to more detailed data or edit specific data. The Add/Edit tab allows the user to add new persons into the application. Any data that is already available in the CDW is preloaded in the application, including contact information, demographic details, tests, symptoms, and other related areas. The Reporting tab displays reports described in more detail subsequently.

4.5.2 Primary care and occupational health workflows

The platform was designed to facilitate multiple workflows from various hospital departments. We started with a results tracking workflow for primary care then expanded functionality for Occupational Health (OH). OH has 3 primary workflows currently supported by the platform: (1) ensure that employees with positive COVID test results are not at work; (2) return to work those who have been out on quarantine

for more than 14 days; and (3) advance those on restricted duties to full duty when clinically appropriate based on guidelines from the United States Centers for Disease Control and Prevention. Using the information captured about the employees such as symptoms, test results, work location, and occupation, a quarantine plan can be conveyed to the employee and supervisor as well as managed on the platform. An algorithm was created to expedite the assignment of a COVID status for an employee based on the results of testing such as "COVID Positive Retest Pending" or "COVID Positive 2^{nd} Negative." A report can be generated by the OH team to show which employees have been out longer than 14 days and have not yet returned to work after 2 negative tests. Employees with restricted duties are captured in the application along with notes from the OH team. Discussions are ongoing about developing workflow features to manage duty restrictions. These features allow the OH team to efficiently allocate which employees need reassessment.

4.5.3 Outbreak investigations report

Our Infection Control staff requested a patient level report that would support them in identifying potential outbreaks at the hospital over a designated time frame. This requires the ability to identify on which day and in which ward location of the hospital a patient first tested positive and any subsequent locations to which the patient was transferred. We configured the report to start 48 hours before illness onset per Centers for Disease Control and Prevention guideline. Because COVIDDMP integrates data from multiple VA facilities within the same interface, it allows for tracking of patients who move across different VA facilities.

COVID Status	Category
Under investigation, monitoring, or pending testing	Employees, Clinical Role Employees, Non-Clinical Role Employees, Quarantined, Undefined Role Employees: Exposure; Return to Duty with Self-Monitoring Patients: Inpatient status
Positive	Employees: COVID Positive, Clinical Role Employees: COVID Positive, Non-Clinical Role Employees: COVID Positive, Undefined Role Patients: Inpatient status PUI/Testing in Progress Patients: Confirmed COVID-19 Positive VABHS at BROCKTON (Inpatient) Patients: Confirmed COVID-19 Positive VABHS at WEST ROXBURY (Inpatient) Patients: Outpatient status Confirmed COVID-19 Positive Patients: COVID-19 Positive Inpatient Deaths
Total tested	Employees: Completed COVID-19 Tests Patients: Completed COVID-19 Tests

Table 4.1: Summary of report aggregation by COVID-19 status, person type, and temporal period

4.5.4 Director's daily briefing report

The VABHS leadership requested a daily briefing classified by clinical staff, patients, and the unique category of employees who are also veteran patients. This report is used to monitor healthcare system capacity, adjust staffing needs, and provide published reports to the healthcare system staff (Table 4.1). Data are provided in aggregate counts of every 24-hour period, as well as cumulatively since March 1, 2020.

4.5.5 Application usage

Since initial deployment the week of March 16, 2020, the application has been used extensively by staff from OH, Infection Control, Primary Care, and other departments. In total, 4629 employees and 7768 patients and have been added for tracking by the application, with 44-759 employees and 128-871 patients newly added each week between March 16 and June 8, 2020 (Figure 4.2). On average, there are 95 $\pm$ 192 updates per day to symptoms, tracking, contact, or other information. Altogether, between April 2 and June 20, 2020, the application was accessed over 46 000 times. In addition to direct usage of the application, data produced by the application are broadly distributed; for example, counts from the Director's Daily Briefing Report (see previous) are distributed to all VABHS employees and affiliates every weekday by email.

4.6 Discussion

Protecting healthcare personnel during the COVID-19 pandemic is critical both for individual employee health and to enable continuity of operations. This is crucially important in cities with high prevalence of the disease.

The data captured by our platform offer an opportunity to create patient-level reports in an effort to support VABHS hospital staff and leadership with COVID-19 monitoring and screening protocols. Integrating data entered by users into the dashboard application with automated data captured by the VA electronic medical record has provided both a historical and real-time view into the operational im-

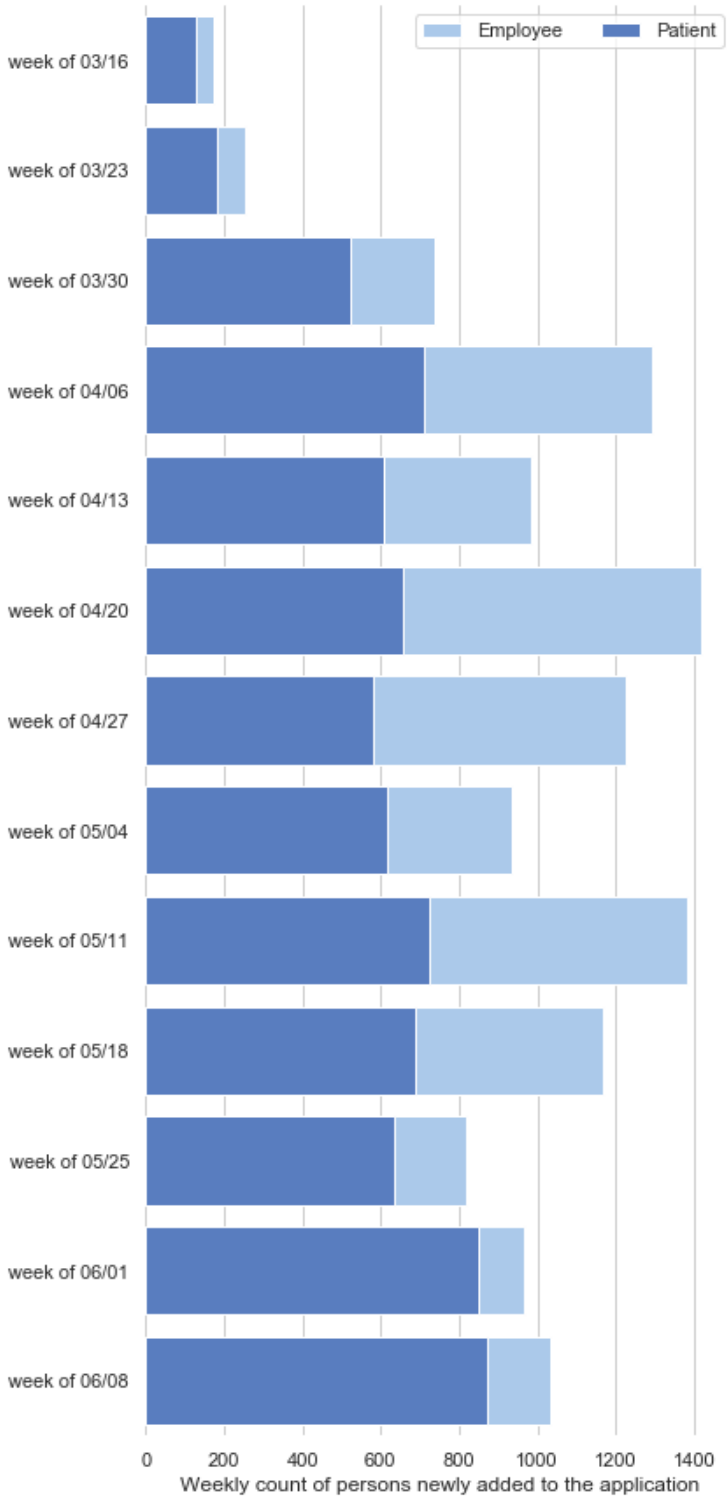


Figure 4.2: Application usage overview. The application has been deployed since the week of March 16, 2020. This figure shows the number of employees and patients newly added for tracking by the application each week between the week of March 16 and the week of June 8, 2020.

pact of COVID-19 at the Boston VA. It has informed our mitigation efforts, and policy decisions on the necessary infrastructure for providing detailed reports for various stakeholders with quick turnaround. Our database contains essential COVID-19 information on patients and employees such as clinical symptoms, known exposures, testing history, testing locations, hospitalizations and death, current patient locations, and quarantined employees. We have begun to use this data to develop aggregate and patient-level reports to support the VA's efforts in identifying new cases, contact tracing, outbreak investigations, and employee workforce management.

On the one hand, the primary limitation of the platform is lack of full integration with the EHR, which necessitates a workaround to address data latency. On the other hand, this lack of integration substantially increases the generalizability of the system to other healthcare systems, as COVIDDMP can be used in any healthcare system after an appropriate script is written to provide input data in the expected format. Further development in FileMan to automate the frequency of data extract will allow additional latency reduction.

4.7 Conclusion

We have developed an informatics solution that facilitates communication and workflow across multiple disciplines to manage hospital employees possibly impacted by the COVID-19 pandemic. Our solution is adaptable to evolving policies and procedures for protecting healthcare personnel as the pandemic progresses, and it is available as open-source software for use by other healthcare systems.

4.8 Author Contributions

NRF*, DCE*, and NVD led application development. NRF, DCE, JL, TCF, F-CS, VN, NL, RZ, SD, SJM, AA, SDG, PC, and SJB implemented the application and supporting infrastructure. RBH, MAP, RSS, DJT, JAD, JMS, SE, BC, MTB, and NVD provided requirements. NRF, DCE, JL, TCF, F-CS, RBH, VN, NL, RZ, SD, SJM, SDG, and NVD drafted the manuscript, and all authors critically reviewed the manuscript. * These authors contributed equally.

4.9 Acknowledgements

The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the U.S. government.

Chapter 5

Sentiment analysis of medical record notes for lung cancer patients at the Department of Veterans Affairs

This Chapter is derived from Elbers et al. (Elbers et al., 2022)

5.1 Abstract

Natural language processing of medical records offers tremendous potential to improve the patient experience. Sentiment analysis of clinical notes has been performed with mixed results, often highlighting the issue that dictionary ratings are not domain specific. Here, for the first time, we re-calibrate the labMT sentiment dictionary on 3.5M clinical notes describing 10,000 patients diagnosed with lung cancer at the Department of Veterans Affairs. The sentiment score of notes was calculated for two years after date of diagnosis and evaluated against a lab test (platelet count) and a combination of data points (treatments). We found that the oncology specific labMT dictionary, after re-calibration for the clinical oncology domain, produces a promising signal in notes that can be detected based on a comparative analysis to the aforementioned parameters.

5.2 Introduction

Estimating the sentiment expressed by text corpora has become popular in the social sciences, especially with the increasing prevalence of social media platforms such as Twitter and Facebook. The Hedonometer, developed by Dodds and Danforth(Dodds & Danforth, 2010)(Dodds et al., 2011), estimates daily global happiness based on Twitter messages (Dodds & Danforth, n.d.), and has been shown to reflect collective attention to global events. The instrument has been successfully utilized in a wide range of domains from quantifying happiness in green spaces (Schwartz, Dodds, O'Neil-Dunne, Danforth, & Ricketts, 2019) to identifying story arcs in palliative

care conversations (Ross et al., 2020) and from understanding social amplification (Alshaabi et al., 2021) to presidential engagement(Minot, Arnold, Alshaabi, Danforth, & Dodds, 2021).

Here, for the first time, we explore the possibility of re-calibrating the Hedonometer sentiment scoring instrument to the clinical oncology domain and utilizing it to identify aspects of the patient trajectory through clinical notes. The resulting signal could be promising as a non-invasive, real-time lens into systems of care, through the interaction between the patient, provider and the healthcare system, making use of all notes entered into the Electronic Health Care Records (EHR). Understanding how care plans and communication are perceived by patients is crucial for continuous improvement of healthcare systems and are currently often quantified only through narrative or survey based studies(Jha, Orav, Zheng, & Epstein, 2008)(Broadbent et al., 2015)(Ruiz-Ceamanos, Spence, & Navarra, 2022).

Sentiment scoring has been explored in the healthcare domain with mixed results. For example, Weismann et al. explored several sentiment scoring methodologies on notes recorded in the ICU (Weissman, Ungar, Harhay, Courtright, & Halpern, 2019) and made available through the MIMIC-III data set(Johnson et al., 2016). They found high variability between methods, but point out a strong association between sentiment and death, that had been seen previously (Waudby-Smith, Tran, Dubin, & Lee, 2018). Weisman et al. suggest that sentiment scoring methodology needs to be more strongly tailored to the healthcare domain and addressed with coverage of specific medical terminology. This argument was highlighted as well by McCoy et al. (McCoy et al., 2015), who performed sentiment scoring on clinical notes utilizing a generic vocabulary scored for polarity (negative vs positive).

We agree that the sentiment associated with medical vocabulary differs greatly from its more common layperson usage, and thus our first step in the present study is to tailor a sentiment dictionary to the oncology domain. For example, the words 'positive' and 'negative' have a very different meaning in a clinical context then they do in generic use. A 'positive' clinical test often signifies an diagnostic indication of a health issue and is almost never a good event in medicine. In this manuscript, we propose and execute a data-driven approach to re-calibration of the original labMT sentiment dictionary, assisted by Subject Matter Experts (SME's) each with 15+ years of experience working in clinical oncology. To foster future research in this area, the re-calibrated score list is made openly available along with this study. Secondly, we evaluated the domain-specific Hedonometer against different oncologic treatments and platelet count lab results. This specific set of parameters was chosen to assess the performance of the Hedonometer, since they are either very objective and often performed, such as the laboratory results known to be indicative of a patient's well being and cancer prognosis (platelet counts)(Sylman et al., 2018)(Zhang, Lv, Yu, & Zhu, 2017)(Maraz et al., 2013), or a combination of medication and date associations that allow for comparison (treatments)(Howlader et al., 2020)(Nabi & Trinh, 2019)(Portier et al., 2013). Our goal is to gauge if the Hedonometer, after re-calibration, can detect a signal indicating clinical trajectory of cancer care using medical record notes.

5.3 Methods

5.3.1 Selected Data

The cohort used in this analysis consists of 10,000 randomly selected patients from the Department of Veterans' Affairs (VA) diagnosed with lung cancer between 2017 and 2019. All notes (3,500,000+) were extracted from the VA's Corporate Data Warehouse (CDW) from date of diagnosis till 2 years after. This resulted in a mean of 886 notes per patient per year and a standard deviation of 2180. To properly validate the signal strength, if detecting any, of the re-calibration of the Hedonometer instrument no sub-selection was made based on note types, i.e. all notes were included.

5.3.2 Re-calibration of the Hedonometer

First, we aimed to revise the LabMT word list for this clinical context, specifically so the sentiment scores were oriented towards the cancer domain. The original study produced the Language Assessment by Mechanical Turk sentiment ratings using an online survey on Amazon's crowdsourced work platform. For the present study, we asked 5 health care providers (3 MDs, 2 nurses) with 10+ years experience in oncology to (re)assess 200 high-importance words (details on their selection will be provided later). In accordance with the original LabMT word list, words are scored on a scale from 1 - 9, with 9 being the happiest or most positive value and 1 being the least positive or most unhappy value. The instructions given to the SME's were the following:

'The results of this survey will be used to measure the happiness of words in the

context of lung oncology care notes. The overall aim is to asses how providers feel about individual words in their clinical context. Please rate the following individual words on a 9 point âunhappy - happyâ scale with 5 being neutral. 1. Read the word. 2. Observe your emotional response. 3. Select score in accordance with your emotional response'.

SME's were able to select a radio button response for each word, with a reminder of score meaning at 1 ('unhappy'), 5 ('neutral'), 9 ('happy').

After exclusion of stop words (Wilbur & Sirotkin, 1992), the list of high-importance words for domain specific re-scoring was selected as follows. We aimed to use both (a) word frequency and (b) the likely difference between layperson and clinical context to identify labMT words most mismatched in sentiment. We chose 100,000 random notes from the larger dataset, and used these notes to design two distinct mechanisms for re-evaluation.

First, we counted all non-labMT words in the notes, and sorted them by frequency. The top 5 words in this lists were 'tab', 'medication', 'prn', 'reviewed' and 'provider'. Second, we parsed the notes for appearances of each anchor labMT word w_i, with labMT sentiment h_i, and identified the adjacent 5 words before and after in the notes. The average sentiment of these neighboring words across all notes was calculated to be h_i^{amb}, the so-called ambient sentiment, and compared with the original labMT rating h_i. Words w_i were then sorted by the magnitude of $h_i^{\delta} = |h_i - h_i^{amb}|$ such that outliers for which the medical context sentiment h_i^{amb} deviated substantially from the context independent sentiment h_i could be identified. Words not scored in the original labMT study such as 'medication', found through frequency ranking, were artificially assigned an h_i^{δ} of 1. To combine these two word lists, we multiplied each

word's frequency in the notes by h_i^{δ}, and truncated after the top 200. The resulting 200 words were deemed most important to re-evaluate, given their prominence in the medical notes and their absent or poor context matching in labMT. Among the 200 words to be scored, a total of 66 new medical domain words were identified, and 134 labMT words were re-scored.

5.3.3 Calculating Sentiment

Sentiment for each note is calculated using the re-calibrated Hedonometer score list, which includes the original words, the original words with new scores, and the new words. In order to focus on the more informative sentiment contributions, words with values between 4 and 6 in the original word list (Dodds et al., 2011) have been excluded from calculating a note's score. Due to the domain specific focus, words scored by the SME's have been included, regardless of score. A note's score is subsequently calculated by, after excluding stop words (Wilbur & Sirotkin, 1992), obtaining the frequency of the unique words occurring in the note, looking up their score in the re-calibrated word list, and multiplying the frequency of the word by its score. After which the sum is taken and the mean for the individual note is calculated (Dodds et al., 2011). Words that are not scored are ignored in all calculations and in line with the Hedonometer strategy no further word cleaning, such as stemming, is performed(Dodds et al., 2011).

Data Type	Start	End	Iteration
Notes	Date of Diagnosis	Date of Diagnosis + 24 months	all
Treatment	Start of Treatment	Start of Treatment + 6 weeks	daily
Platelet Count	Day of Test - 1 week	Day of Test + 1 week	all

Table 5.1: Data Collection

5.3.4 Comparative Analysis

In order to validate the calibration of the Hedonometer instrument, several external parameters were selected to detect a signal in notes. Parameters that were selected for comparison included objective data in the form of lab test outcomes; platelet count, and a combination parameter based on medications, see table 5.1.

Platelet Count

Platelet count was selected as an objective or hard measure, since it has been shown to be an indicator of how well a patient is feeling and has been linked extensively to cancer(Sylman et al., 2018)(Zhang et al., 2017)(Maraz et al., 2013). Additionally, this lab test is ordered quite often, especially in cancer care, varying from every two weeks to daily depending on a patient's state. Since platelet count lab test outcomes come in a multitude of units, data was cleaned up and converted to one unit type ($10^9/L$) before inclusion. For this analysis, platelet count results were divided into three groups: low, normal and high. A low platelet count means below the threshold of 160 $10^9/L$, a normal platelet count is between 160 $10^9/L$ and 410 $10^9/L$, and a high platelet count above 410 $10^9/L$. All clinical notes one week before and one week after a platelet count test was performed were selected and individually scored. These notes were placed in the group in accordance with the lab test outcome. The three groups

were subsequently tested for normality with Shapiro's as well as Anderson's tests. The outcome of these tests determines whether a one-way ANOVA or Kruskal-Wallis test is most appropriate to assess significance. If a significant difference between groups is found, a post-hoc test in the form of Tukey (parametric) or Conover (non-parametric) will be performed to find out which specific between-strata comparisons of platelet count appear to account for the overall difference in sentiment.

Treatments

As a combined data point to validate the re-calibrated Hedonometer against, cancer treatments were chosen. Specifically, chemotherapy, platinum (a form of chemotherapy), targeted and checkpoint therapies were evaluated. First, 39 medications were mapped to the treatments (see table 5.6 in the Appendix), subsequently all these medications and the start of treatment were pulled from the Inpatient, Outpatient, Pharmacy and IV domains in VA's Corporate Data Warehouse (CDW). As treatments might be given in combination or sequence, patients can be counted more then once. Notes were analyzed until six weeks after the start of the specific medication and associated with the treatment. The mean note score for each day was calculated, and a visual inspection of the mean note score fluctuation over time was provided and analyzed with SME's in clinical oncology. Additionally, word shift graphs (Gallagher et al., 2021) were created to gather understanding of which words influenced clear sentiment fluctuations. Lastly, the analysis was ran a second time, utilizing the LabMT dictionary only, to evaluate the difference with the re-calibrated dictionary.

5.4 Results

5.4.1 Re-calibration of the Hedonometer

Please see 5.8 Appendix for a complete list of the 200 words selected based on their surrounding sentiment and frequency to be re-scored by SME's. Figure 5.1 shows that our selection of 200 words will cover 30% of note text.

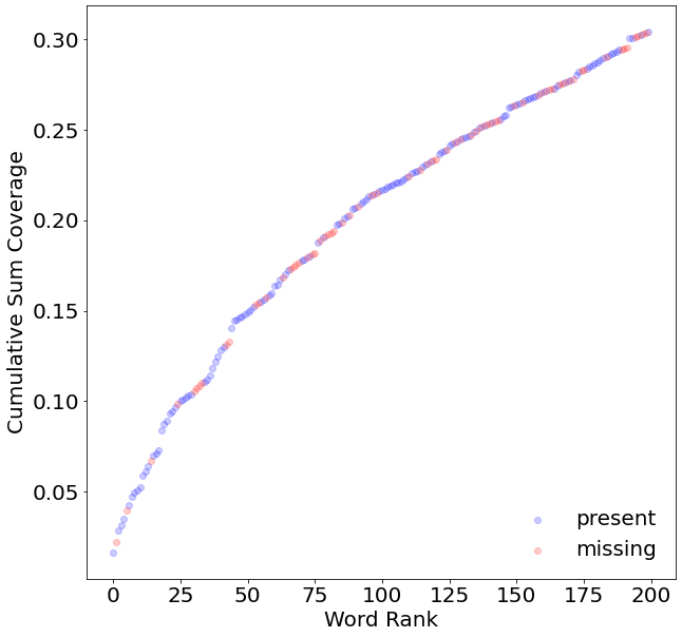


Figure 5.1: Words are ranked based on the product of their clinical note coverage and the difference in word score to a word's ambient sentiment score.

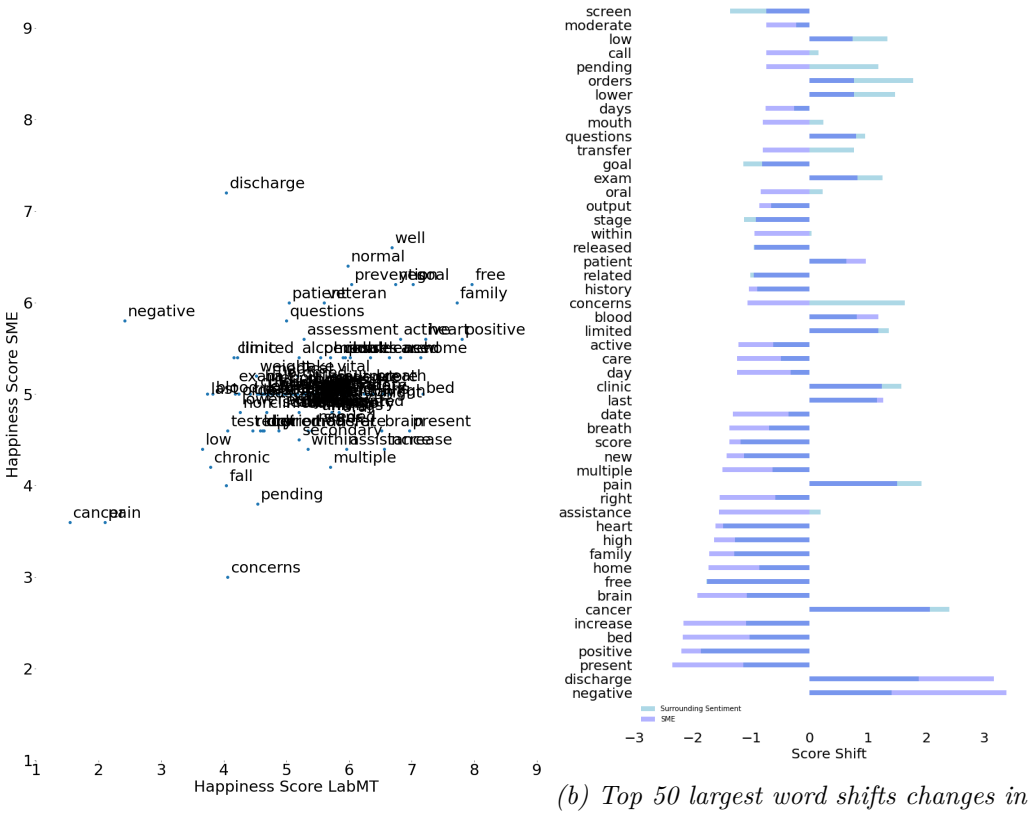
Mean standard deviation of re-scoring the high-importance words between the SME's was 0.51 on the 9-point scale. The mean of re-scored words, 4.93, is lower then the mean of the original labMT dataset, namely 5.37. Full scores and standard deviation are available in table 5.4 in the Appendix. The addition and updates of the oncology domain specific words result in a word list of 10,253, an increase of 66

words from the original 10,187. Examination of the re-scored words, see figure 5.2a, shows that word 'positive' is scored less high by the SME's then in LabMT (5.6 vs. 7.8), while the word 'negative' is scored much higher (5.8 vs. 2.4). Other words standing out include 'discharge', 'pain', 'cancer', 'veteran' and 'family'. To compare, figure 5.2b shows the top 50 words with the largest shifts in word score for both the SME's evaluation and based on the surrounding sentiment calculation. A subset of random clinical notes was evaluated for confirmatory no's; standard questions generally answered with the word 'no'. Questions similar to 'have you recently travelled outside of the country?', with answers of 'no' were identified, however did not occur in high frequency. It appears that repetitive questions are often answered with the word 'negative', e.g. 'negative for fever, chills, blurring of vision, redness of eye, nausea' a word that was re-scaled by the SME's. The word 'no' itself is ranked 410, and was thus not highly influential on score calculation.

5.4.2 Comparative Analysis

Platelet Count

Every group with platelet count results, low, normal and high, was comprised evenly to contain 39882 unique notes. This number was equivalent to the group with the lowest count of notes (high group), for the larger groups a random subset was selected to create equal groups. All three groups failed both the Shapiro and Anderson normality tests. Although visual inspection of QQ-plots in addition to histograms did appear to come close to normality, the high number of samples might have played a factor in failing the tests. To be on the safe side, it was decided to test non-parametric, thus



(a) LabMT vs SME word scores

(b) Top 50 largest word shifts changes in respect to LabMT

Figure 5.2: Word score shifts due to calibration by SME's. Figure 5.2a on the left compares the LabMT scores to the scores assigned by the SME's. The right figure 5.2b compares the ambient sentiment for each anchor term and the eventual assessment by the SME's.

a Kruskal-Wallis was subsequently performed. The result of the Kruskal-Wallis test showed a significant difference ($p = 2.79 \times 10^{-06}$) and post-hoc Conover results are displayed in table 5.2.

	High	Low	Normal
High	1	0.56	$3.49e^{-06}$
Low	0.56	1	$4.84e^{-05}$
Normal	$3.49e^{-06}$	$4.84e^{-05}$	1

Table 5.2: Platelet Count - Post Hoc Conover Test

A significant difference in note score is present between the high platelet count group and the normal platelet count group, as well as between the low platelet count group and the normal platelet count group. However no difference was found between the high and low platelet count group. The high platelet count group had a median note score of 5.308; mean of 5.355, the low platelet count group a median note score of 5.316; mean of 5.356, while the normal platelet count group had a slightly higher median note score of 5.325; mean of 5.363.

Treatments

Evaluating the scoring of daily notes generated during the six weeks after the start of either one of the four different treatments; chemotherapy (633160 notes), platinum (317392 notes), checkpoint (230500 notes) and targeted (139908 notes), shows a cyclical weekly pattern, see figure 5.3. Day 40 for targeted therapy has the lowest count of data points, namely 1980 notes.

According to SME's in clinical lung cancer, this cyclical pattern can be explained due to patients' weekly visits with their providers at which point they are asked about side-effects, bringing down the sentiment score of generated notes. Day 21

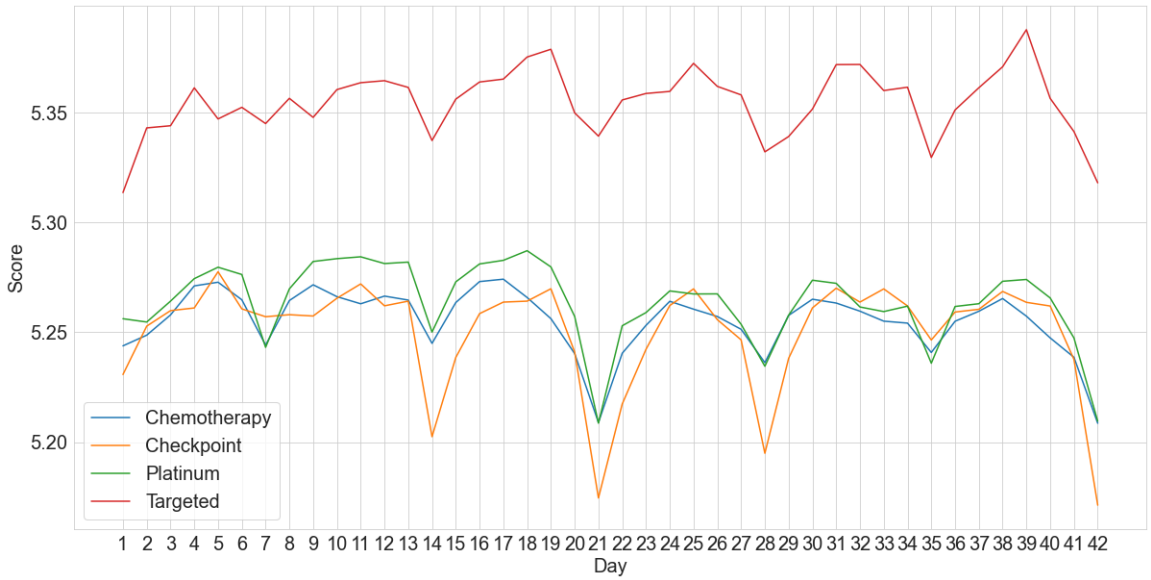
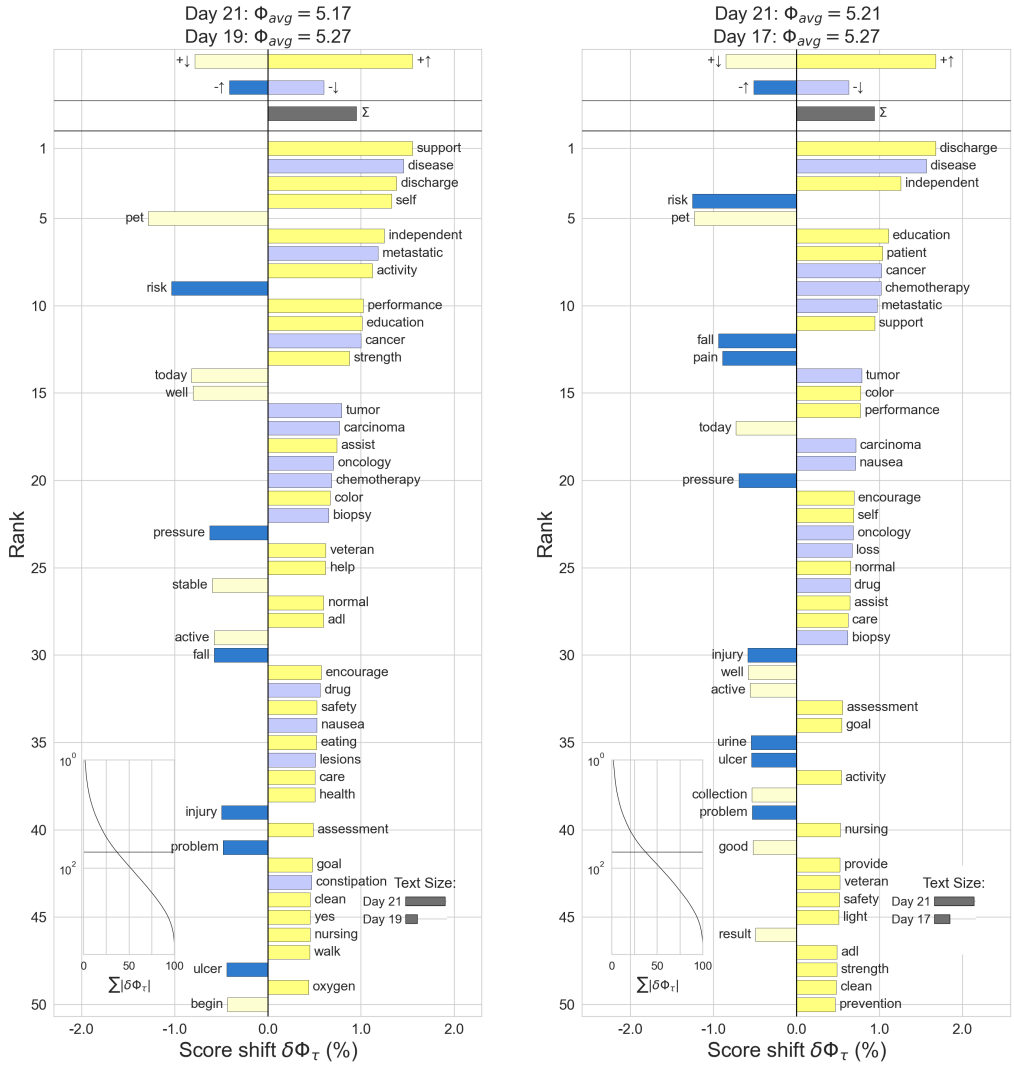


Figure 5.3: Note scores on a day to day basis per treatment arm, starting at the day of treatment till six weeks from date of start of treatment.

after treatment is known to be related to the worst side effects, after which side effects begin improving. Targeted therapy tends to produce fewer side-effects and is often given to patients in better health, which could explain the better overall score and reduced fluctuation. However, targeted therapy still appears to also follow the cyclical pattern present in the other three treatments. Reasons behind checkpoint therapy having the deepest dip in sentiment score, as opposed to chemotherapy, could not be identified by our SME's.

To better identify what creates the dip in note scores on day 21, word shift plots are created, see figure 5.4. Words related to lung cancer treatment such as 'lung', 'cancer', 'treatment', 'dose', 'mouth' and 'chemotherapy' appear to drive the score down. While more generic words, such as 'care', 'support', 'activity', 'patient' and 'independent', related to patient care seem to positively influence the note scores on peak days. Notably the text size is larger on day 21, associating more volume of note



(a) Checkpoint therapy arm *(b) Chemotherapy arm*

Figure 5.4: Using notes authored on Day 21 of treatment as a reference, word-shift graphs detail the words influencing the drop in sentiment when compared with day 19 (left) and and 17 (right). Looking at the comparison between days 19 and 21 on the left, words appearing on the left side of the graph contribute positively to day 21, while words on the right side contribute positively to day 19 (there are many more of this type). For example, the relatively positive words 'support', 'discharge', and 'independent' are more common on day 19. The relatively negative words 'disease' and 'metastatic' are less common on day 19. Going against the overall trend, the relatively positive words 'today', 'well', and 'stable' are more common on day 21. The relatively negative words 'risk', 'pressure', and 'fall' are less common on day 21.

generation during visits. Rerunning the analysis with the original LabMT dictionary showed a similar cyclical pattern, however the overall note scores were 0.3 points higher. This difference was attributed to the prevalence of missing clinical words in LabMT, which were generally scored low by the SME's.

5.5 Discussion

We have designed and implemented a data-driven method for re-calibration of an existing sentiment scoring instrument, the Hedonometer, in a specific healthcare domain. This re-calibration has been proposed numerous times in previous research on sentiment in clinical notes (Weissman et al., 2019)(Waudby-Smith et al., 2018)(McCoy et al., 2015) as a solution to the mixed results found when utilizing existing vocabularies in this context. However, to the authors' awareness, re-calibration has not been done before. The re-calibration of the labMT sentiment dictionary for the oncology healthcare domain resulted in 200 re-scored words, which together cover 30% of clinical note text for the oncology domain.

The successful application of the domain-specific Hedonometer in our comparative analysis suggests that a signal is present and measurable in clinical notes. Validating the instrument on laboratory test outcomes, we found a significant difference in normal versus low or high platelet counts. Additionally, a consistent cyclic effect is visible when scoring notes in relation to cancer treatment timelines, showing a comparable cycle for all treatments. This cycle can be explained by weekly provider visits discussing side effects and treatment itself, with a clear dip on day 21 for the chemotherapy (incl. platinum) and checkpoint treatments indicating the side-effects

being at their worst. Discussion with clinical SME's underscores this belief. When examining score shift with day 21, treatment related words indeed appear to drive the note score down. This negative shift is also consistent with Portier et al. (Portier et al., 2013), who perform sentiment analysis on online cancer support forums and find that side-effects of treatments are a topic with a clear negative sentiment shift. Though an explanation for the low dip in checkpoint treatment specifically was not presented by the SME's, it can be hypothesized that this treatment often includes further advanced cancer cases or cases in palliative care. It would be interesting to see if splitting out durvalumab, often given to less advanced cases, would generate higher sentiment.

Our evaluation of the domain-specific Hedonometer is far from exhaustive, and future research should be done to understand better how and where the instrument can be utilized and what its limitations are in the oncology and healthcare domain. The finding that a signal appears present, detectable, and consistent is encouraging and the authors hypothesize next research steps as stratification of the cohort and researching difference between note types, both of which have not been addressed here, and evaluation against other parameters. Other parameters could include, for example, the the VA - Frailty Index(Cheng et al., 2021)(DuMontier et al., 2021), a well-validated performance measurement based on a large set of variables across different healthcare domains, or the VA-CAN score(Osborne, Veigulis, Arreola, Roosli, & Curtin, 2020), surgery or neutrophil counts. If the signal remains robust, clinical and research implementation can vary from quality control of a new medication or treatment plan, to better understanding certain subgroups to quantifying differences between hospital, provider and nurse care.

5.6 Acknowledgements

The authors like to thank Dr. Albert Lin, Dr. David Tuck, Karen Murray and Karen Visnaw for their help in re-calibration, and Dr. Albert Lin and Dr. Mikaela Fudolig for their input and suggestions. The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the U.S. government.

5.7 Data Availability

Patient related and note data cannot be shared publicly because it involves sensitive human subject data. Data may be available for researchers who meet the criteria for access to confidential data after evaluation from VA Research and Development Committees. As a VA national legal policy (VHA Directive 1605.01), VA will only share patient data if there is a fully executed contract in place for the specific project. A common contractual mechanism utilized for this type of sharing is a 'Cooperative Research and Development' (CRADA) agreement. These contracts are typically negotiated in collaboration with VA national Office of General Council (OGC) and lawyers from the collaborating institution. These national sharing policies and standards also apply to deidentified data. In addition, if a contract is in place allowing sharing of deidentified data outside of VA, then VA national policy (VHA Directive 1605.01), states that deidentification certification needs to be met by Expert Determination. The expert determination requires independent assessment from an experienced master or PhD in biostatistics, from a third party not involved in the project, and may

require outside funding to support. In addition, for an outside entity to preform research on VA patient data, IRB as well as VA Research and Development Committee approval is required for the specific project. Data requests may be sent to: VA Information Resource Center (VIReC) Building 18 Hines VA Hospital (151V) 5000 S. 5th Avenue Hines, IL 60141-3030 708-202-2413 708-202-2415 (fax) virec@va.gov.

5.8 Appendix

rnk	word	coverage	diff.	diff x cov
0	patient	0.016321	0.488	0.007959
1	tab	0.006045	1	0.006045
2	take	0.006038	0.886	0.005348
3	lung	0.003023	1.574	0.004757
4	tablet	0.003708	1.264	0.004687
5	medications	0.004553	1	0.004553
6	non	0.00275	1.566	0.004305
7	one	0.005072	0.791	0.00401
8	history	0.002204	1.607	0.003542
9	pulse	0.001321	2.347	0.003101
10	discharge	0.001599	1.879	0.003004
11	mouth	0.006201	0.478	0.002964
12	chest	0.00215	1.343	0.002887
13	left	0.003353	0.844	0.002829
14	medication	0.002824	1	0.002824
15	treatment	0.002768	0.919	0.002544
16	score	0.001502	1.635	0.002455
17	scale	0.001412	1.558	0.002201
18	active	0.011381	0.192	0.002186
19	assessment	0.003469	0.568	0.00197
20	clinic	0.001325	1.425	0.001889
21	yes	0.003985	0.464	0.001847
22	heart	0.001277	1.412	0.001804
23	daily	0.002788	0.646	0.0018
24	prn	0.001711	1	0.001711
25	lower	0.001281	1.315	0.001685
26	signs	0.001072	1.533	0.001643
27	screen	0.000641	2.553	0.001636
28	new	0.001461	1.115	0.001629
29	stage	0.000748	2.137	0.001598
30	reviewed	0.001555	1	0.001555
31	provider	0.00151	1	0.00151
32	denies	0.001481	1	0.001481
33	refills	0.001473	1	0.001473
34	chronic	0.000946	1.531	0.001448
35	staff	0.001169	1.231	0.001438
36	bed	0.002312	0.614	0.001419
37	veteran	0.003716	0.37	0.001376
38	every	0.003759	0.365	0.001373
39	days	0.002686	0.51	0.001371
40	right	0.003551	0.384	0.001365

Table 5.3: Top 40 words based on rank of Surrounding Sentiment ∗ Text Coverage

Words	Score	SD	Words	Score	SD	Words	Score	SD
patient	6	2	oral	5	0	ulcer	3.6	1.02
tab	5	0	hcl	4.6	0.8	note	5	0
take	5.2	0.4	qty	5	0	albuterol	4.2	0.98
lung	5	0	intake	5	0	past	5	0
tablet	5	0	mobility	5.2	0.4	biopsy	3.2	1.47
medications	4.6	0.49	respiratory	5	0	concerns	3	1.26
non	4.8	0.4	back	5	0	catheter	5	0
one	5	0	icd	4.6	0.8	precautions	5.2	0.4
history	4.8	0.4	mild	5.4	0.49	sodium	5	0
pulse	5.4	0.8	inj	4.8	0.4	bilateral	5	0
discharge	7.2	0.98	inhl	5	0	oncology	3.4	1.36
mouth	5	0	day	5	0	increase	4.4	1.2
chest	5	0	dressing	5	0	home	5.4	0.49
left	5	0	medical	5	0	transfer	4.6	0.8
medication	4.8	0.4	imaging	5	0	date	5	0
treatment	5	0.63	allergies	5	1.26	neck	5	0
score	5	0	sct	4.6	0.8	expiration	4.4	1.2
scale	5	0	intact	6	1.26	free	6.2	1.17
active	5.6	0.8	use	5	0	vital	5.2	0.4
assessment	5.6	1.2	diet	4.6	0.8	ref	4.8	0.4
clinic	5.4	0.8	inpatient	3.6	1.36	post	5	0
yes	6.2	1.6	per	5	0	screening	5	0
heart	5.6	1.2	reports	5	0	evidence	5	0
daily	5	0	labs	5	0	limited	5.4	0.8
prn	5	0	last	5	0	minutes	5	0
lower	5	0	questions	5.8	1.17	inhale	5.4	0.8
signs	5	0	chemotherapy	4.4	0.8	need	5	0
screen	5	0	twice	5	0	expr	4.6	0.8
new	5.4	1.02	units	5	0	secondary	4.5	0.87
stage	4.8	1.6	cancer	3.6	1.5	copd	3.8	1.17
reviewed	5.2	0.4	family	6	1.26	adl	5.8	0.98
provider	5.4	0.49	soln	5	0	brain	4.6	1.96
denies	4.2	0.75	orders	5	0	prior	4.6	0.49
refills	5.2	0.4	edema	3.8	1.17	issu	4.4	0.8
chronic	4.2	0.98	breath	5	0	verified	4.8	0.4
staff	5	0	required	4.6	0.8	clinical	4.8	0.4
bed	5	0	positive	5.6	0.8	abdomen	5	0
veteran	6	2	goal	6.2	0.75	male	5.4	0.8
every	5	0	number	5	0	scan	5	0
days	5	0	released	5.4	0.8	call	5	0

Table 5.4: Re-scoring outcomes from SME's, part 1

Words	Score	SD	Words	Score	SD	Words	Score	SD
right	5	0	alcohol	5.4	1.02	within	4.4	1.2
procedure	5	0	lock	4.6	0.8	carcinoma	2.6	1.62
outpatient	4.8	0.4	recent	5	0	nodule	3.8	1.47
lobe	5.2	0.98	unit	5	0	year	5	0
pain	3.6	2.15	report	5	0	two	5	0
skin	5	0	lymph	5.2	0.4	ordered	5	0
dry	4.6	0.8	weight	5.2	1.6	assistance	4.4	0.8
high	5	0	occur	5	0	prevention	6.2	1.17
related	5	0	months	5	0	negative	5.8	1.6
needs	5	0	caregiver	6.6	1.02	rate	5	0
test	4.6	0.8	level	5	0	puffs	4.4	1.2
cap	5	0	well	6.6	1.36	low	4.4	1.2
dose	5	0	capsule	5	0	room	5	0
interventions	5	0	exam	5	0	findings	5	0
sig	4.8	0.4	moisture	5	0	moderate	4.6	0.8
half	5	0	consult	4.8	0.4	person	5.4	0.8
present	4.6	0.8	needed	5	0	constipation	2.8	1.83
nutrition	5.6	0.8	shift	5	0	mmol	5	0
location	5	1.67	output	5	0	wbc	5	0
current	5	0	tablets	5	0	care	5.4	0.8
blood	5	0	may	5	0	multiple	4.2	1.6
cell	5	0	name	5	0	extremities	5	0
normal	6.4	1.5	glucose	5.2	0.4	shortness	4	1.26
bowel	5	0	pending	3.8	1.6	results	5.4	0.8
fall	4	0.89	nausea	3.4	1.36	metastatic	2.8	1.6
instructions	5	0	chloride	5	0	meals	5.6	0.8
bid	5	0	four	4.8	0.4			

Table 5.5: Re-scoring outcomes from SME's, part 2

Generic	category	Brand
abraxane	chemotherapy	
afatinib	targeted	gilotrif
alectinib	targeted	alecensa
atezolizumab	checkpoint	tecentriq
bevacizumab	targeted	avastin
brigatinib	targeted	alunbrig
capmatinib	targeted	
carboplatin	platinum	
ceritinib	targeted	zykadia
cisplatin	platinum	
crizotinib	targeted	xalkori
dabrafenib	targeted	tafinlar
dacomitinib	targeted	vizimpro
docetaxel	chemotherapy	taxotere
doxorubicin	chemotherapy	adriamycin
durvalumab	checkpoint	imfinzi
entrectinib	targeted	
erlotinib	targeted	
etoposide	chemotherapy	etopophos
etoposide	chemotherapy	
everolimus	targeted	afinitor
gefitinib	targeted	iressa
gemcitabine	chemotherapy	gemzar
ipilimumab	checkpoint	yervoy
larotrectinib	targeted	vitrakvi
lorlatinib	targeted	lorbrena
mechlorethamine	chemotherapy	mustargen
necitumumab	targeted	portrazza
nivolumab	checkpoint	opdivo
osimertinib	targeted	tagrisso
paclitaxel	chemotherapy	taxol
pembrolizumab	checkpoint	keytruda
pemetrexed	chemotherapy	alimta
ramucirumab	targeted	cyramza
selpercatinib	targeted	
topotecan	chemotherapy	hycamtin
trametinib	targeted	mekinist
vinorelbine	chemotherapy	navelbine

Table 5.6: Medication treatment matrix

References

I. o. M. (2003). *Key Capabilities of an Electronic Health Record System*. National Academies Press (US). Retrieved 2022-02-10, from https://www.ncbi.nlm.nih.gov/books/NBK221802/ doi: 10.17226/10781

V. H. A. (n.d.). *About VHA - Veterans Health Administration* [General Information]. Retrieved 2022-02-11, from https://www.va.gov/health/aboutvha.asp (Accepted: 20210423)

Activiti. (2022, February). Activiti. Retrieved 2022-02-11, from https://github.com/Activiti/Activiti (original-date: 2012-09-13T11:34:43Z)

Adalja, A. A., Toner, E., & Inglesby, T. V. (2020, April). Priorities for the US Health Community Responding to COVID-19. *JAMA*, *323*(14), 1343–1344. Retrieved 2022-02-13, from https://doi.org/10.1001/jama.2020.3413 doi: 10.1001/jama.2020.3413

Adams, J. G., & Walls, R. M. (2020, April). Supporting the Health Care Workforce During the COVID-19 Global Epidemic. *JAMA*, *323*(15), 1439–1440. Retrieved 2022-02-13, from https://doi.org/10.1001/jama.2020.3972 doi: 10.1001/jama.2020.3972

Alba, P. R., Gao, A., Lee, K. M., Anglin-Foote, T., Robison, B., Katsoulakis, E., ... Lynch, J. A. (2021, December). Ascertainment of Veterans With Metastatic Prostate Cancer in Electronic Health Records: Demonstrating the Case for Natural Language Processing. *JCO Clinical Cancer Informatics*(5), 1005–1014. Retrieved 2022-02-11, from https://ascopubs.org/doi/abs/10.1200/CCI.21.00030 (Publisher: Wolters Kluwer) doi: 10.1200/CCI.21.00030

Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2021). The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009â2020. *EPJ data science*, *10*(1), 15. (Publisher: Springer Berlin Heidelberg)

Atherton, J. (2011, March). Development of the Electronic Health Record. *AMA Journal of Ethics*, *13*(3), 186–189. Retrieved 2022-02-10, from https://journalofethics.ama-assn.org/article/

development-electronic-health-record/2011-03 (Publisher: American Medical Association) doi: 10.1001/virtualmentor.2011.13.3.mhst1-1103

Baker, M. G., Peckham, T. K., & Seixas, N. S. (2020, April). Estimating the burden of United States workers exposed to infection or disease: A key factor in containing risk of COVID-19 infection. *PLOS ONE*, *15*(4), e0232452. Retrieved 2022-02-13, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232452 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0232452

Barnett, G., Justice, N., Somand, M., Adams, J., Waxman, B., Beaman, P., ... Greenlie, J. (1979, September). COSTAR, a computer-based medical information system for ambulatory care. *Proceedings of the IEEE*, *67*(9), 1226–1237. (Conference Name: Proceedings of the IEEE) doi: 10.1109/PROC.1979.11438

Bates, D. W., Teich, J. M., Lee, J., Seger, D., Kuperman, G. J., Ma'Luf, N., ... Leape, L. (1999). The Impact of Computerized Physician Order Entry on Medication Error Prevention. *Journal of the American Medical Informatics Association : JAMIA*, *6*(4), 313–321. Retrieved 2022-02-10, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC61372/

Ben-Israel, D., Jacobs, W. B., Casha, S., Lang, S., Ryu, W. H. A., de Lotbiniere-Bassett, M., & Cadotte, D. W. (2020, March). The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine*, *103*, 101785. Retrieved 2022-02-11, from https://www.sciencedirect.com/science/article/pii/S0933365719303951 doi: 10.1016/j.artmed.2019.101785

Berry, I., Soucy, J.-P. R., Tuite, A., & Fisman, D. (2020, April). Open access epidemiologic data and an interactive dashboard to monitor the COVID-19 outbreak in Canada. *CMAJ*, *192*(15), E420–E420. Retrieved 2022-02-13, from https://www.cmaj.ca/content/192/15/E420 (Publisher: CMAJ Section: Letters) doi: 10.1503/cmaj.75262

Blue Ribbon Panel Report. (2016). , 74.

Broadbent, E., Wilkes, C., Koschwanez, H., Weinman, J., Norton, S., & Petrie, K. J. (2015, November). A systematic review and meta-analysis of the Brief Illness Perception Questionnaire. *Psychology & Health*, *30*(11), 1361–1385. Retrieved 2022-02-25, from https://doi.org/10.1080/08870446.2015.1070851 (Publisher: Routledge _eprint: https://doi.org/10.1080/08870446.2015.1070851) doi: 10.1080/08870446.2015.1070851

Brown, S. (2003, March). VistA, a U.S. Department of Veterans Affairs national-scale HIS. *International Journal of Medical Informatics*, *69*(2-3), 135–156. Retrieved 2022-02-10, from https://linkinghub.elsevier.com/retrieve/pii/S1386505602001314 doi: 10.1016/S1386-5056(02)00131-4

Brunette, C. A., Miller, S. J., Majahalme, N., Hau, C., MacMullen, L., Advani, S., ... Vassy, J. L. (2020). Pragmatic Trials in Genomic Medicine: The Integrating Pharmacogenetics In Clinical Care (I-PICC) Study. *Clinical and Translational Science*, *13*(2), 381–390. Retrieved 2022-02-12, from https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12723 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12723) doi: 10.1111/cts.12723

Budrionis, A., & Bellika, J. G. (2016, December). The Learning Healthcare System: Where are we now? A systematic review. *Journal of Biomedical Informatics*, *64*, 87–92. Retrieved 2022-02-26, from https://www.sciencedirect.com/science/article/pii/S1532046416301319 doi: 10.1016/j.jbi.2016.09.018

Burnette, E. (n.d.). *Eclipse IDE Pocket Guide.* Retrieved 2022-02-11, from https://www.oreilly.com/library/view/eclipse-ide-pocket/0596100655/ (ISBN: 9780596100650)

Burton, C., Elliott, A., Cochran, A., & Love, T. (2018, September). Do healthcare services behave as complex systems? Analysis of patterns of attendance and implications for service delivery. *BMC Medicine*, *16*(1), 138. Retrieved 2022-02-11, from https://doi.org/10.1186/s12916-018-1132-5 doi: 10.1186/s12916-018-1132-5

Casey, J. A., Schwartz, B. S., Stewart, W. F., & Adler, N. E. (2016, March). Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health*, *37*(1), 61–81. Retrieved 2022-02-10, from https://www.annualreviews.org/doi/10.1146/annurev-publhealth-032315-021353 doi: 10.1146/annurev-publhealth-032315-021353

CDC COVID-19 Response Team. (2020, April). Characteristics of Health Care Personnel with COVID-19 - United States, February 12-April 9, 2020. *MMWR. Morbidity and mortality weekly report*, *69*(15), 477–481. doi: 10.15585/mmwr.mm6915e6

Cheng, D., DuMontier, C., Yildirim, C., Charest, B., Hawley, C. E., Zhuo, M., ... Orkaby, A. R. (2021, July). Updating and Validating the U.S. Veterans Affairs Frailty Index: Transitioning From ICD-9 to ICD-10. *The Journals of Gerontology: Series A*, *76*(7), 1318–1325. Retrieved 2022-02-25, from https://doi.org/10.1093/gerona/glab071 doi: 10.1093/gerona/glab071

Cheng, D., Ramos-Cejudo, J., Tuck, D., Elbers, D., Brophy, M., Do, N., & Fillmore, N. (2019, August). External validation of a prognostic model for mortality among patients with nonâsmall-cell lung cancer using the Veterans Precision Oncology Data Commons. *Seminars in Oncology*, *46*(4), 327–333. Retrieved 2022-02-13, from https://www.sciencedirect.com/science/article/pii/S0093775419300557 doi: 10.1053/j.seminoncol.2019

.09.006

Chou, R., Dana, T., Buckley, D. I., Selph, S., Fu, R., & Totten, A. M. (2020, July). Epidemiology of and Risk Factors for Coronavirus Infection in Health Care Workers. *Annals of Internal Medicine*, *173*(2), 120–136. Retrieved 2022-02-13, from https://www.acpjournals.org/doi/full/10.7326/M20-1632 (Publisher: American College of Physicians) doi: 10.7326/M20-1632

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., ... Prior, F. (2013, December). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, *26*(6), 1045–1057. Retrieved 2022-02-12, from https://doi.org/10.1007/s10278-013-9622-7 doi: 10.1007/s10278-013-9622-7

Corporate Data Warehouse (CDW). (n.d.). Retrieved 2022-02-12, from https://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm

Dankar, F. K., Ptitsyn, A., & Dankar, S. K. (2018, April). The development of large-scale de-identified biomedical databases in the age of genomicsâprinciples and challenges. *Human Genomics*, *12*(1), 19. Retrieved 2022-02-12, from https://doi.org/10.1186/s40246-018-0147-5 doi: 10.1186/s40246-018-0147-5

D'Avolio, L., Ferguson, R., Goryachev, S., Woods, P., Sabin, T., O'Neil, J., ... Fiore, L. (2012, June). Implementation of the Department of Veterans Affairs' first point-of-care clinical trial. *Journal of the American Medical Informatics Association*, *19*(e1), e170–e176. Retrieved 2022-02-12, from https://doi.org/10.1136/amiajnl-2011-000623 doi: 10.1136/amiajnl-2011-000623

Deep Dive: How a Health Tech Sprint Pioneered an AI Ecosystem. (2019, February). Retrieved 2022-02-13, from https://digital.gov/2019/02/27/how-a-health-tech-sprint-inspired-an-ai-ecosystem/

Dhond, R., Elbers, D., Majahalme, N., Dipietro, S., Goryachev, S., Acher, R., ... Do, N. V. (2021, July). ProjectFlow: a configurable workflow management application for point of care research. *JAMIA Open*, *4*(3), ooab074. Retrieved 2022-02-11, from https://doi.org/10.1093/jamiaopen/ooab074 doi: 10.1093/jamiaopen/ooab074

Do, N., Grossman, R., Feldman, T., Fillmore, N., Elbers, D., Tuck, D., ... Brophy, M. (2019, August). The Veterans Precision Oncology Data Commons: Transforming VA data into a national resource for research in precision oncology. *Seminars in Oncology*, *46*(4), 314–320. Retrieved 2022-02-11, from https://www.sciencedirect.com/science/article/pii/S0093775419300521 doi: 10.1053/j.seminoncol.2019.09.002

Dodds, P. S., & Danforth, C. M. (n.d.). *Hedonometer.* Retrieved 2022-01-28, from https://hedonometer.org

Dodds, P. S., & Danforth, C. M. (2010, August). Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness*

Studies, *11*(4), 441–456. Retrieved 2021-08-19, from http://link.springer.com/10.1007/s10902-009-9150-9 doi: 10.1007/s10902-009-9150-9

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011, December). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. Retrieved 2021-08-19, from http://arxiv.org/abs/1101.5120 (arXiv: 1101.5120) doi: 10.1371/journal.pone.0026752

Dong, E., Du, H., & Gardner, L. (2020, May). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. Retrieved 2022-02-13, from https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext (Publisher: Elsevier) doi: 10.1016/S1473-3099(20)30120-1

DuMontier, C., Fillmore, N. R., Yildirim, C., Cheng, D., La, J., Orkaby, A. R., ... Driver, J. A. (2021, January). Contemporary Analysis of Electronic Frailty Measurement in Older Adults with Multiple Myeloma Treated in the National US Veterans Affairs Healthcare System. *Cancers*, *13*(12), 3053. Retrieved 2022-02-25, from https://www.mdpi.com/2072-6694/13/12/3053 (Number: 12 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/cancers13123053

Ehrlich, H., McKenney, M., & Elkbuli, A. (2020, July). Protecting our healthcare workers during the COVID-19 pandemic. *The American Journal of Emergency Medicine*, *38*(7), 1527–1528. Retrieved 2022-02-13, from https://www.ajemjournal.com/article/S0735-6757(20)30252-7/fulltext (Publisher: Elsevier) doi: 10.1016/j.ajem.2020.04.024

Elbers, D. C., Fillmore, N. R., Sung, F.-C., Ganas, S. S., Prokhorenkov, A., Meyer, C., ... Do, N. V. (2020, September). The Veterans Affairs Precision Oncology Data Repository, a Clinical, Genomic, and Imaging Research Database. *Patterns*, *1*(6), 100083. Retrieved 2022-02-11, from https://www.sciencedirect.com/science/article/pii/S2666389920301112 doi: 10.1016/j.patter.2020.100083

Elbers, D. C., La, J., Minot, J., Gramling, R. E., Brophy, M. T., Vo, N., ... Danforth, C. M. (2022). Sentiment analysis of medical notes for lung cancer patients in the department of veterans affairs. *In Submission*.

Ernst, M. E., & Lund, B. C. (2010). Renewed Interest in Chlorthalidone: Evidence From the Veterans Health Administration. *The Journal of Clinical Hypertension*, *12*(12), 927–934. Retrieved 2022-02-11, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-7176.2010.00373.x (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-7176.2010.00373.x) doi: 10.1111/j.1751-7176.2010.00373.x

FACT SHEET: At Cancer Moonshot Summit, Vice President Biden An-

nounces New Actions to Accelerate Progress Toward Ending Cancer As We Know It. (2016, June). Retrieved 2022-02-12, from https://obamawhitehouse.archives.gov/the-press-office/2016/06/28/fact-sheet-cancer-moonshot-summit-vice-president-biden-announces-new

Fihn, S. D., Francis, J., Clancy, C., Nielson, C., Nelson, K., Rumsfeld, J., ... Graham, G. L. (2014, July). Insights From Advanced Analytics At The Veterans Health Administration. *Health Affairs*, *33*(7), 1203–1211. Retrieved 2022-02-13, from https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2014.0054 (Publisher: Health Affairs) doi: 10.1377/hlthaff.2014.0054

Fillmore, N., Ramos-Cejudo, J., Cheng, D., Tuck, D. P., Sheikh, A. R., Chen, D., ... Do, N. (2019, May). A predictive model for survival in non-small cell lung cancer (NSCLC) based on electronic health record (EHR) and tumor sequencing data at the Department of Veterans Affairs (VA). *Journal of Clinical Oncology*, *37*(15_suppl), 109–109. Retrieved 2022-02-13, from https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.109 (Publisher: Wolters Kluwer) doi: 10.1200/JCO.2019.37.15_suppl.109

Fillmore, N. R., Elbers, D. C., La, J., Feldman, T. C., Sung, F.-C., Hall, R. B., ... Do, N. V. (2020, November). An application to support COVID-19 occupational health and patient tracking at a Veterans Affairs medical center. *Journal of the American Medical Informatics Association*, *27*(11), 1716–1720. Retrieved 2022-02-13, from https://doi.org/10.1093/jamia/ocaa162 doi: 10.1093/jamia/ocaa162

Fiore, L., Ferguson, R. E., Brophy, M., Kudesia, V., Shannon, C., Zimolzak, A., ... Lavori, P. (2016, February). Implementation of a Precision Oncology Program as an Exemplar of a Learning Health Care System in the VA. *Federal Practitioner*, *33*(Suppl 1), 26S–30S. Retrieved 2022-02-11, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6375409/

Fiore, L. D., Brophy, M., Ferguson, R. E., DâAvolio, L., Hermos, J. A., Lew, R. A., ... Lavori, P. W. (2011, April). A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clinical Trials*, *8*(2), 183–195. Retrieved 2022-02-12, from https://doi.org/10.1177/1740774511398368 (Publisher: SAGE Publications) doi: 10.1177/1740774511398368

Fiore, L. D., Brophy, M. T., Turek, S., Kudesia, V., Ramnath, N., Shannon, C., ... Lavori, P. (2016, January). The VA Point-of-Care Precision Oncology Program: Balancing Access with Rapid Learning in Molecular Cancer Medicine. *Biomarkers in Cancer*, *8*, BIC.S37548. Retrieved 2022-02-12, from https://doi.org/10.4137/BIC.S37548 (Publisher: SAGE Publications Ltd STM) doi: 10.4137/BIC.S37548

Friedman, C. P., Rubin, J. C., & Sullivan, K. J. (2017, August). Toward an Information Infrastructure for Global Health Improvement. *Yearbook of Medical Informatics*, *26*(1), 16–23. Retrieved 2022-02-10, from http://www.thieme-connect.de/DOI/DOI?10.15265/IY-2017-004 (Publisher: Georg Thieme Verlag KG) doi: 10.15265/IY-2017-004

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2021, December). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, *10*(1), 4. Retrieved 2021-08-26, from https://epjds.epj.org/articles/epjdata/abs/2021/01/13688_2021_Article_260/13688_2021_Article_260.html (Number: 1 Publisher: Springer Berlin Heidelberg) doi: 10.1140/epjds/s13688-021-00260-3

Gan, W. H., Lim, J. W., & Koh, D. (2020, June). Preventing Intra-hospital Infection and Transmission of Coronavirus Disease 2019 in Health-care Workers. *Safety and Health at Work*, *11*(2), 241–243. Retrieved 2022-02-13, from https://www.sciencedirect.com/science/article/pii/S209379112030161X doi: 10.1016/j.shaw.2020.03.001

Genomic Data Commons. (n.d.). Retrieved 2022-02-12, from https://portal.gdc.cancer.gov/projects/VAREPOP-APOLLO

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020, May). A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings*, *2020*, 191–200. Retrieved 2022-02-11, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/

Godderis, L., Boone, A., & Bakusic, J. (2020, July). COVID-19: a new work-related disease threatening healthcare workers. *Occupational Medicine*, *70*(5), 315–316. Retrieved 2022-02-13, from https://doi.org/10.1093/occmed/kqaa056 doi: 10.1093/occmed/kqaa056

Golinelli, D., Boetto, E., Carullo, G., Nuzzolese, A. G., Landini, M. P., & Fantini, M. P. (2020, November). Adoption of Digital Technologies in Health Care During the COVID-19 Pandemic: Systematic Review of Early Scientific Literature. *Journal of Medical Internet Research*, *22*(11), e22280. Retrieved 2022-02-13, from https://www.jmir.org/2020/11/e22280 (Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/22280

Greenhalgh, T., & Papoutsi, C. (2018, June). Studying complexity in health services research: desperately seeking an overdue paradigm shift. *BMC Medicine*, *16*(1), 95. Retrieved 2022-02-11, from https://doi.org/10.1186/s12916

-018-1089-4 doi: 10.1186/s12916-018-1089-4

Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. "O'Reilly Media, Inc.". (Google-Books-ID: cVlPDwAAQBAJ)

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016, September). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. Retrieved 2022-02-13, from https://doi.org/10.1056/NEJMp1607591 (Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp1607591) doi: 10.1056/NEJMp1607591

Hakkoum, H., Abnane, I., & Idri, A. (2022, March). Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, *117*, 108391. Retrieved 2022-02-11, from https://www.sciencedirect.com/science/article/pii/S1568494621011522 doi: 10.1016/j.asoc.2021.108391

HealthITAnalytics. (2019, December). *VA Announces New National Artificial Intelligence Institute.* Retrieved 2022-02-11, from https://healthitanalytics.com/news/va-announces-new-national-artificial-intelligence-institute

Hersh, W. R. (2007). Adding Value to the Electronic Health Record Through Secondary Use of Data for Quality Assurance, Research, and Surveillance. , 2.

Hollander, J. E., & Carr, B. G. (2020, April). Virtually Perfect? Telemedicine for Covid-19. *New England Journal of Medicine*, *382*(18), 1679–1681. Retrieved 2022-02-13, from https://doi.org/10.1056/NEJMp2003539 (Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp2003539) doi: 10.1056/NEJMp2003539

Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., ... Feuer, E. J. (2020, August). The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *New England Journal of Medicine*, *383*(7), 640–649. Retrieved 2022-02-25, from https://doi.org/10.1056/NEJMoa1916623 (Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1916623) doi: 10.1056/NEJMoa1916623

Jha, A. K., Orav, E. J., Zheng, J., & Epstein, A. M. (2008, October). Patients' Perception of Hospital Care in the United States. *New England Journal of Medicine*, *359*(18), 1921–1931. Retrieved 2022-02-25, from https://doi.org/10.1056/NEJMsa0804116 (Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsa0804116) doi: 10.1056/NEJMsa0804116

Jodogne, S. (2018, June). The Orthanc Ecosystem for Medical Imaging. *Journal of Digital Imaging*, *31*(3), 341–352. Retrieved 2022-02-13, from https://doi.org/10.1007/s10278-018-0082-y doi: 10.1007/s10278-018-0082-y

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi,

M., ... Mark, R. G. (2016, May). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035. Retrieved 2022-02-12, from https://www.nature.com/articles/sdata201635 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/sdata.2016.35

Judd, C. M., Nusairat, J. F., & Shingler, J. (2008). Beginning Groovy and Grails. , 76.

Judson, T. J., Odisho, A. Y., Neinstein, A. B., Chao, J., Williams, A., Miller, C., ... Gonzales, R. (2020, June). Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. *Journal of the American Medical Informatics Association*, *27*(6), 860–866. Retrieved 2022-02-13, from https://doi.org/10.1093/jamia/ocaa051 doi: 10.1093/jamia/ocaa051

Kelley, M. J., Duffy, J., Hintze, B. J., Williams, C. D., & Spector, N. L. (2017, May). Implementation of precision oncology in the Veterans Health Administration (VHA). *Journal of Clinical Oncology*, *35*(15_suppl), 6507–6507. Retrieved 2022-02-12, from https://ascopubs.org/doi/abs/10.1200/JCO.2017.35.15_suppl.6507 (Publisher: Wolters Kluwer) doi: 10.1200/JCO.2017.35.15_suppl.6507

Ko, R. K., Lee, S. S., & Wah Lee, E. (2009, January). Business process management (BPM) standards: a survey. *Business Process Management Journal*, *15*(5), 744–791. Retrieved 2022-02-11, from https://doi.org/10.1108/14637150910987937 (Publisher: Emerald Group Publishing Limited) doi: 10.1108/14637150910987937

Langer, S. G., Tellis, W., Carr, C., Daly, M., Erickson, B. J., Mendelson, D., ... Zhu, W. (2015, February). The RSNA Image Sharing Network. *Journal of Digital Imaging*, *28*(1), 53–61. Retrieved 2022-02-12, from https://doi.org/10.1007/s10278-014-9714-z doi: 10.1007/s10278-014-9714-z

Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., & Zwarenstein, M. (2015, May). The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*, *350*, h2147. Retrieved 2022-02-11, from https://www.bmj.com/content/350/bmj.h2147 (Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting) doi: 10.1136/bmj.h2147

Maraz, A., Furak, J., Varga, Z., Kahan, Z., Tiszlavicz, L., & Hideghety, K. (2013, April). Thrombocytosis Has a Negative Prognostic Value in Lung Cancer. *Anticancer Research*, *33*(4), 1725–1729. Retrieved 2022-02-25, from https://ar.iiarjournals.org/content/33/4/1725 (Publisher: International Institute of Anticancer Research Section: Clinical Studies)

McCoy, T. H., Castro, V. M., Cagan, A., Roberson, A. M., Kohane, I. S., & Perlis, R. H. (2015, August). Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Elec-

tronic Health Record Study. *PLOS ONE*, *10*(8), e0136341. Retrieved 2022-01-28, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136341 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0136341

McLachlan, S., Potts, H. W. W., Dube, K., Buchanan, D., Lean, S., Gallagher, T., ... Fenton, N. (2018, April). The Heimdall framework for supporting characterisation of learning health systems. *BMJ Health & Care Informatics*, *25*(2), 77–87. Retrieved 2022-02-11, from https://informatics.bmj.com/lookup/doi/10.14236/jhi.v25i2.996 doi: 10.14236/jhi.v25i2.996

Meng, F., Morioka, C. A., & Elbers, D. C. (2019, September). Generating Information Extraction Patterns from Overlapping and Variable Length Annotations using Sequence Alignment. *arXiv:1908.03594 [cs]*. Retrieved 2022-02-11, from http://arxiv.org/abs/1908.03594 (arXiv: 1908.03594)

Minot, J. R., Arnold, M. V., Alshaabi, T., Danforth, C. M., & Dodds, P. S. (2021, April). Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. *PLOS ONE*, *16*(4), e0248880. Retrieved 2022-02-25, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0248880 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0248880

Morris, Z. S., Wooding, S., & Grant, J. (2011, December). The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, *104*(12), 510–520. Retrieved 2022-02-11, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3241518/ doi: 10.1258/jrsm.2011.110180

Mustra, M., Delac, K., & Grgic, M. (2008, September). Overview of the DICOM standard. In *2008 50th International Symposium ELMAR* (Vol. 1, pp. 39–44). (ISSN: 1334-2630)

Nabi, J., & Trinh, Q.-D. (2019). *New Cancer Therapies Are Great-But Are They Helping Everyone? | Health Affairs Forefront.* Retrieved 2022-02-25, from https://www.healthaffairs.org/do/10.1377/forefront.20190410.590278/full/

Nagesh, S., & Chakraborty, S. (2020, June). Saving the frontline health workforce amidst the COVID-19 crisis: Challenges and recommendations. *Journal of Global Health*, *10*(1), 010345. doi: 10.7189/jogh-10-010345

Nourani, A., Ayatollahi, H., & Dodaran, M. S. (2019, September). Clinical Trial Data Management Software: A Review of the Technical Features. *Reviews on Recent Clinical Trials*, *14*(3), 160–172. doi: 10.2174/1574887114666190207151500

Office for Civil Rights, O. (2012, September). *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy*

Rule [Text]. Retrieved 2022-02-12, from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (Last Modified: 2022-02-02T18:35:41-0500)

Olsen, L., Aisner, D., & McGinnis, J. M. (2007). *The Learning Healthcare System*. National Academies Press (US). Retrieved 2022-02-11, from https://www.ncbi.nlm.nih.gov/books/NBK53494/ doi: 10.17226/11903

ORD VHA Directive, Handbooks, and Program Guides – 1200 series. (2022, February). Retrieved 2022-02-12, from https://www.research.va.gov/resources/policies/handbooks.cfm

Osborne, T. F., Veigulis, Z. P., Arreola, D. M., Roosli, E., & Curtin, C. M. (2020, July). Automated EHR score to predict COVID-19 outcomes at US Department of Veterans Affairs. *PLOS ONE*, *15*(7), e0236554. Retrieved 2022-02-25, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236554 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0236554

Patel, P. D., Cobb, J., Wright, D., Turer, R. W., Jordan, T., Humphrey, A., ... Rosenbloom, S. T. (2020, July). Rapid development of telehealth capabilities within pediatric patient portal infrastructure for COVID-19 care: barriers, solutions, results. *Journal of the American Medical Informatics Association*, *27*(7), 1116–1120. Retrieved 2022-02-13, from https://doi.org/10.1093/jamia/ocaa065 doi: 10.1093/jamia/ocaa065

Point of Care Research (POC-R). (n.d.). Retrieved 2022-02-11, from https://www.research.va.gov/programs/csp/point-of-care.cfm

Portier, K., Greer, G. E., Rokach, L., Ofek, N., Wang, Y., Biyani, P., ... Yen, J. (2013, December). Understanding Topics and Sentiment in an Online Cancer Survivor Community. *JNCI Monographs*, *2013*(47), 195–198. Retrieved 2022-02-25, from https://doi.org/10.1093/jncimonographs/lgt025 doi: 10.1093/jncimonographs/lgt025

Reeves, J. J., Hollandsworth, H. M., Torriani, F. J., Taplitz, R., Abeles, S., Tai-Seale, M., ... Longhurst, C. A. (2020, June). Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *Journal of the American Medical Informatics Association*, *27*(6), 853–859. Retrieved 2022-02-13, from https://doi.org/10.1093/jamia/ocaa037 doi: 10.1093/jamia/ocaa037

Ross, L., Danforth, C. M., Eppstein, M. J., Clarfeld, L. A., Durieux, B. N., Gramling, C. J., ... Gramling, R. (2020, April). Story Arcs in Serious Illness: Natural Language Processing features of Palliative Care Conversations. *Patient Education and Counseling*, *103*(4), 826–832. Retrieved 2022-03-07, from https://www.sciencedirect.com/science/article/pii/S0738399119305282 doi: 10.1016/j.pec.2019.11.021

Ruiz-Ceamanos, A., Spence, C., & Navarra, J. (2022, February). Individual Differences in Chemosensory Perception Amongst Cancer Patients Undergoing Chemotherapy: A Narrative Review. *Nutrition and Cancer*, *0*(0), 1–15. Retrieved 2022-02-25, from https://doi.org/10.1080/01635581.2021.2000625 (Publisher: Routledge _eprint: https://doi.org/10.1080/01635581.2021.2000625) doi: 10.1080/01635581.2021.2000625

Schwartz, A. J., Dodds, P. S., O'Neil-Dunne, J. P. M., Danforth, C. M., & Ricketts, T. H. (2019, December). Visitors to urban greenspace have higher sentiment and lower negativity on Twitter. *People and Nature*, *1*(4), 476–485. Retrieved 2022-02-25, from https://onlinelibrary.wiley.com/doi/10.1002/pan3.10045 doi: 10.1002/pan3.10045

Semple, S., & Cherrie, J. W. (2020, June). Covid-19: Protecting Worker Health. *Annals of Work Exposures and Health*, *64*(5), 461–464. Retrieved 2022-02-13, from https://doi.org/10.1093/annweh/wxaa033 doi: 10.1093/annweh/wxaa033

Shanafelt, T., Ripp, J., & Trockel, M. (2020, June). Understanding and Addressing Sources of Anxiety Among Health Care Professionals During the COVID-19 Pandemic. *JAMA*, *323*(21), 2133–2134. Retrieved 2022-02-13, from https://doi.org/10.1001/jama.2020.5893 doi: 10.1001/jama.2020.5893

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019, May). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, *7*(2), e12239. Retrieved 2022-02-11, from https://medinform.jmir.org/2019/2/e12239 (Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/12239

Singer, D. S., Jacks, T., & Jaffee, E. (2016, September). A U.S. "Cancer Moonshot" to accelerate cancer research. *Science*, *353*(6304), 1105–1106. Retrieved 2022-02-12, from https://www.science.org/doi/full/10.1126/science.aai7862 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.aai7862

Staa, T.-P. v., Goldacre, B., Gulliford, M., Cassell, J., Pirmohamed, M., Taweel, A., ... Smeeth, L. (2012, February). Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ*, *344*, e55. Retrieved 2022-02-11, from https://www.bmj.com/content/344/bmj.e55 (Publisher: British Medical Journal Publishing Group Section: Analysis) doi: 10.1136/bmj.e55

Sylman, J. L., Boyce, H. B., Mitrugno, A., Tormoen, G. W., Thomas, I.-C., Wagner, T. H., ... Mallick, P. (2018, April). A Temporal Examination of

Platelet Counts as a Predictor of Prognosis in Lung, Prostate, and Colon Cancer Patients. *Scientific Reports*, *8*(1), 6564. Retrieved 2022-02-25, from https://www.nature.com/articles/s41598-018-25019-1 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41598-018-25019-1

TCIA Collections. (n.d.). Retrieved 2022-02-13, from https://www.cancerimagingarchive.net/collections/

TOP Health Sprint. (n.d.). Retrieved 2022-02-13, from https://tophealth.pif.gov/

VA FileMan Technical Manual. (2021). , 98.

Vaishya, R., Haleem, A., Vaish, A., & Javaid, M. (2020, July). Emerging Technologies to Combat the COVID-19 Pandemic. *Journal of Clinical and Experimental Hepatology*, *10*(4), 409–411. Retrieved 2022-02-13, from https://www.jcehepatology.com/article/S0973-6883(20)30061-X/fulltext (Publisher: Elsevier) doi: 10.1016/j.jceh.2020.04.019

van der Laan, A. L., & Boenink, M. (2015, March). Beyond Bench and Bedside: Disentangling the Concept of Translational Research. *Health Care Analysis*, *23*(1), 32–49. Retrieved 2022-02-11, from https://doi.org/10.1007/s10728-012-0236-x doi: 10.1007/s10728-012-0236-x

The Veterans Precision Oncology Data Commons. (n.d.). Retrieved 2022-02-13, from https://vpodc.data-commons.org

Vickers, A. J., & Scardino, P. T. (2009, March). The clinically-integrated randomized trial: proposed novel method for conducting large trials at low cost. *Trials*, *10*(1), 14. Retrieved 2022-02-11, from https://doi.org/10.1186/1745-6215-10-14 doi: 10.1186/1745-6215-10-14

Waudby-Smith, I. E. R., Tran, N., Dubin, J. A., & Lee, J. (2018, June). Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLOS ONE*, *13*(6), e0198687. Retrieved 2022-01-28, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198687 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0198687

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... Stuart, J. M. (2013, October). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. Retrieved 2022-02-12, from https://www.nature.com/articles/ng.2764 (Number: 10 Publisher: Nature Publishing Group) doi: 10.1038/ng.2764

Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R., & Halpern, S. D. (2019, January). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics*, *89*, 114–121. Retrieved 2021-08-19, from https://www.sciencedirect.com/science/article/pii/S1532046418302284 doi:

10.1016/j.jbi.2018.12.001

Wen, H.-C., Ho, Y.-S., Jian, W.-S., Li, H.-C., & Hsu, Y.-H. E. (2007, May). Scientific production of electronic health record research, 1991â2005. *Computer Methods and Programs in Biomedicine*, *86*(2), 191–196. Retrieved 2022-02-10, from https://linkinghub.elsevier.com/retrieve/pii/S0169260707000314 doi: 10.1016/j.cmpb.2007.02.002

White, S. A., & Bock, C. (2011). *BPMN 2.0 Handbook Second Edition: Methods, Concepts, Case Studies and Standards in Business Process Management Notation*. Future Strategies Inc. (Google-Books-ID: 9U3DO5PoTDQC)

Wilbur, W. J., & Sirotkin, K. (1992, February). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55. Retrieved 2022-02-04, from https://doi.org/10.1177/016555159201800106 (Publisher: SAGE Publications Ltd) doi: 10.1177/016555159201800106

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016, March). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. Retrieved 2022-03-06, from https://www.nature.com/articles/sdata201618 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/sdata.2016.18

Wissel, B. D., Van Camp, P. J., Kouril, M., Weis, C., Glauser, T. A., White, P. S., ... Dexheimer, J. W. (2020, July). An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time. *Journal of the American Medical Informatics Association*, *27*(7), 1121–1125. Retrieved 2022-02-13, from https://doi.org/10.1093/jamia/ocaa071 doi: 10.1093/jamia/ocaa071

Wouters, R. H., van der Graaf, R., Voest, E. E., & Bredenoord, A. L. (2020). Learning health care systems: Highly needed but challenging. *Learning Health Systems*, *4*(3), e10211. Retrieved 2022-02-11, from https://onlinelibrary.wiley.com/doi/abs/10.1002/lrh2.10211 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lrh2.10211) doi: 10.1002/lrh2.10211

Zhang, X., Lv, Z., Yu, H., & Zhu, J. (2017, June). The clinicopathological and prognostic role of thrombocytosis in patients with cancer: A meta-analysis. *Oncology Letters*, *13*(6), 5002–5008. Retrieved 2022-02-25, from https://www.spandidos-publications.com/10.3892/ol.2017.6054 (Publisher: Spandidos Publications) doi: 10.3892/ol.2017.6054

Zullig, L. L., Jackson, G. L., Dorn, R. A., Provenzale, D. T., McNeil, R., Thomas, C. M., & Kelley, M. J. (2012, June). Cancer Incidence Among Patients of the U.S. Veterans Affairs Health Care System. *Military Medicine*, *177*(6), 693–701. Retrieved 2022-02-12, from https://doi.org/10.7205/MILMED-D-11-00434 doi: 10.7205/MILMED-D-11-00434