Applying Network Analysis and Supervised Learning to Electronic Clinical Notes to Improve Operational Suicide Risk Prevention at an Academic Medical Center

By

Kevin Jason Krause

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University in

partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December 17, 2022

Approved:

Colin Walsh, M.D.

Sharon Davis, Ph.D.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3ST. *3 Step Theory*
ANOVA. *Analysis of Variance*
AP. *Average Precision*
AUPRC. *Area Under the Precision Recall Curve*
AUROC. *Area Under the Receiver Operating Characteristic*
CDC. *Centers for Disease Control*
CDM. *Common Data Model*
CDS. *Clinical Decision Support*
CPM. *Counts per Million*
CSSRS. *Columbia Suicide-Severity Rating Scale*
CUI. *Concept Unique Identifier*
DNN. *Deep Neural Network*
ECT. *Electroconvulsive Therapy*
ED. *Emergency Department*
EHR. *Electronic Health Record*
HIV. *Human Immunodeficiency Virus*
HL7. *Health Level 7*
i2b2. *Informatics for Integrating Biology and the Bedside*
IMV. *Integrated Motivational-Volitional*
IRB. *Institutional Review Board*
ITSB. *Interpersonal Theory of Suicidal Behavior*
LR. *Logistic Regression*
MDI. *Mean Decrease in Impurity*
NLP. *Natural Language Processing*
OHDSI. *Observational Health Data Sciences and Informatics*
OMOP. *Observational Medical Outcomes Partnership*
OSA. *Obstructive Sleep Apnea*
PPV. *Positive Predictive Value*
PTSD. *Posttraumatic Stress Disorder*
RFA. *Risk Factor Analysis*
SCS. *Suicide Crisis Syndrome*
SLM. *Supervised Learning Model*
SMOTE. *Synthetic Minority Oversampling Technique*
TBI. *Traumatic Brain Injury*
TFIDF. *Total Frequency Inverse Document Frequency*
UMLS. *Unified Medical Language System*
VUMC. *Vanderbilt University Medical Center*

Chapter 1: Introduction

Suicide is a long-standing public health concern and modern informatics methods are creating new opportunities to understand suicide and prevent harm. In this thesis, I will explore suicide risk factors in clinical notes and use clinical notes to improve an operational suicide risk prediction model.

Suicide

Statistics

In 2019, suicide was the 10th leading cause of death in the United States.[1] Approximately 800,000 people died from suicide between 1999 and 2019, and the suicide rate has been steadily increasing for the past decade.[1] For the past 50 years, the suicide rate has increased, despite most other major causes of mortality decreasing in rate.[2] The National Institute of Health (NIH) reports that funding awards with the term "suicide" in the title have increased from 22.4M in 2001, to 199.4M in 2011 and 466.7M in 2021.[3] Despite massive national investments and increases in public awareness, suicide is still a significant threat to public health.

History

Early English writings on suicide date back to the 17th century, when clergyman John Sym wrote the oldest extant English text on suicide, entitled *Life's Preservative Against Self-Killing*.[4,5] Sym's work, and the work of others in this time period discussed suicide through a theological and moral lens.[5] In the later 18th century, researchers began to discuss the causes of suicide.[5] In 1790, clergyman Charles Moore noted an intergenerational trend in suicidal tendency, possibly the first written hint of a genetic suicide risk factor.[5,6] Moore noted correlations between suicide and alcoholism, gender, and nationality; further, he remarked on the challenges related to obtaining accurate suicide data.[5,6] In the early 19th century, researchers began to discuss suicide as a sociological illness, paving the way toward the modern medical understanding of suicidality.[5] The 19th century saw a boom of written research and discussion of suicidal causes, trends, and treatments, culminating with Emile Durkheim's work *Suicide: A Study in Sociology* which is often credited with laying the foundation for modern western suicidology.[7,8]

Theories and Causes

Durkheim posits that suicide arises from dysregulation of social integration and/or moral regulation. Durkheim's four hypothesized types of suicide represent imbalances of these forces: egoistic (too little integration), altruistic (too much integration), anomic (too little regulation), and fatalistic (too much regulation).[7,8] Leading modern suicidology frameworks borrow heavily from Durkheim's emphasis on social integration[7]; these include Joiner's Interpersonal Theory of Suicidal Behavior (ITSB)[9,10], Klonsky's Three-Step-Theory (3ST)[11–13], and O'Connor's Integrated Motivational-Volitional (IMV) model[14] (Figure 1).

The ITSB hypothesizes that suicidal desire emerges from "thwarted belongingness and perceived burdensomeness" and progresses to suicidal behavior via acquired capability from repeated traumatic experiences.[9,10] Similarly, the 3ST hypothesizes that pain and hopelessness cause suicidal ideation, which then escalates in the absence of connectedness via acquired suicidal capability, leading to suicide attempt.[11–13] The works of Durkheim[8], Joiner[10], Klonsky[11], and O'Connor[14] all agree that suicidal ideation is a preceding step to suicidal behavior and social factors can protect against suicidal progression. Further, Joiner[9,10], Klonsky[11–13], and O'Connor[14] all emphasize the importance of *acquired capability to commit suicide* within the suicidal process.

Treatment and Screening

Despite the seriousness and persistence of suicidal ideation, suicidal persons usually experience suicidal intent and capability in brief, intense episodes of crisis.[15,16] Thus, support-givers, educators, and medical professionals can provide support to suicidal persons by helping manage capability, intent, and means. The most effective tools for managing and surviving suicidal crises include physician education, lethal means restriction, and psychiatric treatment.[17,18] The centers for disease control (CDC) also cites evidence that cognitive behavioral therapy and dialectical behavioral therapy can reduce suicide risk.[19–21]

Clinicians are increasingly focused on improving screening for suicidal ideation and behavior to improve intervention rate for at-risk persons. Structured approaches are better at assessing suicide risk than routine clinical interviews.[22] One validated structured approach for assessing suicide risk is the Columbia Suicide-Severity Rating Scale (CSSRS), quantifies suicidality in terms of four constructs by means of a questionnaire.[22] Though effective, questionnaires rely on clinical action to assess risk. More recently, clinical informaticians are developing automated tools for assessing suicide-risk with existing data from patients' electronic health records (EHRs).[23]

The Integrated Motivational-Volitional Model of Suicidality

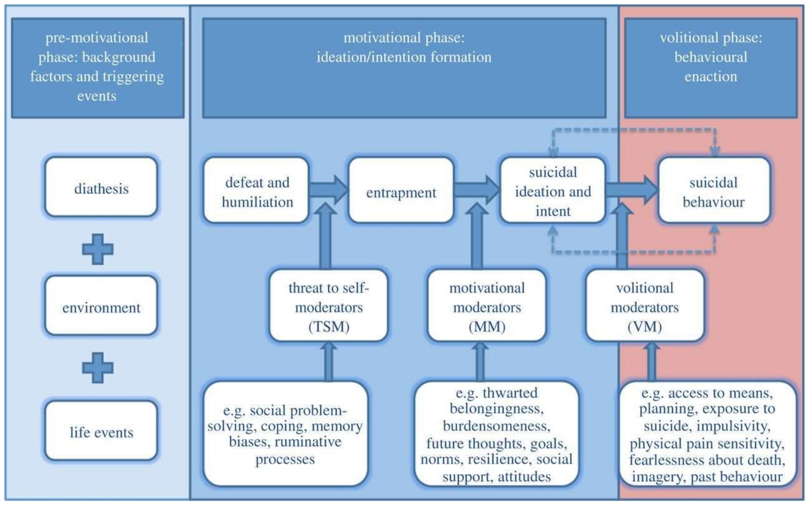


Figure-1: This figure depicts the idea-to-action framework described by O'Connor and Kirtley in *The integrated motivational–volitional model of suicidal behaviour.*[14]

Informatics

Prior Work

Prominent suicidology-informatics research topics include analyses to ascertain suicidal phenotypes, identify risk factors for suicidal behavior, and predict risk of suicidal behavior. Suicidal phenotype ascertainment refers to classifying patients/encounters as positive or negative for various suicidal traits/events. Further, ascertainment efforts may also attempt to date-stamp event phenotypes (e.g., a suicide attempt). For instance, suicidal phenotype ascertainment could refer to classifying patients as suicidal ideators/non-ideators or encounters as suicide attempt events/non-events. Suicidal phenotype ascertainment is an important preceding step to supervised analyses, which require suicidal outcome-labelled datasets. Researchers Bejan and Walsh at Vanderbilt University Medical Center (VUMC) demonstrate the effectiveness of using billing codes to extract suicidal ideation and suicide attempt phenotypes from the EHR.[24] Other researchers demonstrate alternative ascertainment methods, such as the use of natural language processing (NLP).[25]

Suicide risk factor analysis (RFA) refers to quantifying associations between patient-measures and suicidal outcomes. Patient-measures include demographics, lifestyle factors, and past medical histories. Previous suicide RFAs identify depression, hopelessness, childhood sexual abuse, alcohol use disorder, and gender and sexual identity as associated risk factors for suicide attempt.[26–28] Epidemiologists are divided over whether associations (i.e., RFA results) can establish causality.[29,30] Some favor probabilistic models, which support the use of associations to establish causality, while others favor deterministic models based in pure science causal modelling.[29,30] Recent trends favor counterfactually-based probabilistic models, a blend of both approaches, which aims to more faithfully support the aims of epidemiology.[29] Causal implications aside, risk analyses can help identify at-risk individuals (e.g., with relative risk) and develop predictive risk models (e.g., with feature selection).[23,25,31–35]

Network analysis is another method which has been used to explore suicide risk factors. In 2017, De Beurs described the potential merits of applying network analysis to suicidology, which had not been done prior.[36] De Beurs' theoretical approach uses a suicidal behavior questionnaire based on the integrated motivational-volitional model of suicidality[14] to generate networks with node strength defined by question response agreement magnitude and edge strength defined by intra-individual question response associations.[36] Two studies[37,38] on veterans and one study[39] on adolescents used psychological assessments and questionnaires to generate association networks to explore how different symptoms and factors relate to suicidal ideation. Bloch-Elkouby et. al. used network analysis to validate a newly proposed condition, Suicide Crisis Syndrome (SCS).[40] The authors showed that the symptoms of SCS were highly clustered together and centered around entrapment, supporting the proposed diagnostic criteria for the condition.[40] Beyond these, few other studies have used network analysis to examine suicidal behavior.

Suicide risk prediction refers to quantifying risk of future suicidal behaviors (e.g., suicide attempt or suicidal ideation). Risk prediction models are commonly based on statistics[31], machine learning[23], scientific knowledge[22], or a combination thereof. Statistical and knowledge-based risk models exhibit superior clinical familiarity and explainability than machine learning models. Further, machine learning model development can be more demanding, particularly with model calibration and validation. Despite these limitations, machine learning models are increasingly popular due to superior achievable predictive performances. Researchers Walsh, Ribeiro, and Franklin demonstrate the effectiveness of machine learning-based risk prediction for both suicidal ideation and suicide attempt with the development of VSAIL, the Vanderbilt Suicide Attempt and Ideation Likelihood prediction model.[23] Despite validation challenges, VSAIL achieves clinically meaningful performance in operation at VUMC[41], underscoring the promise of machine learning based suicide risk prediction models. Future work on VSAIL is expected to be allocated towards model-retraining, continuous-learning, and implementation/decision support.

Structured and Unstructured Data

Clinical informaticians source data from the EHR, which contains both *structured* and *unstructured* data. Structured data are those which are organized into consistent, often tabular, structures, and unstructured data are those which have no formal consistency or organization.[42] Structured data include medication lists, billing codes, and visit occurrences, while unstructured data include written clinical assessments, patient portal messages, and radiology reports. Organizations create data-structuring standards to support information exchange, compliance, and usability within and between enterprises. For example, Health-Level 7 (HL7) International develops the consolidated Clinical Document Architecture to standardize health-information messaging, helping facilitate health-information exchanges and clinical decision support (CDS) applications.[43] Observational Health Data Sciences and Informatics (OHDSI) develops the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to support interoperable observational data modelling.[44] The national Institute of Health Roadmap National Centers for Biomedical Computing sponsors the Informatics for Integrating Biology and the Bedside (i2b2) model, which provides a platform for integrating standard EHR data with clinical research and genomic data.[45] Further, computational analyses (e.g., machine learning) often use strict dimensional specifications and mappings which require structured data as inputs.

Though structured data are more convenient to use, a massive amount of information is stored as unstructured data in the EHR. On the higher end, researchers estimate that as much 95% of big-data is unstructured.[46] Evidence shows that unstructured data are more beneficial than structured data for machine learning tasks.[25,35,47] Further, suicide analysis is limited by a lack structured indicators for suicidal phenotypes in the EHR[47], and suicidal phenotype ascertainment can be improved by analyzing unstructured data.[48] These results indicate that unstructured data contain powerful embedded signals which may help suicide-researchers improve their analyses.

Natural Language Processing

Researchers commonly use natural language processing (NLP) tools to parse, extract, and structure information from unstructured data to complement and/or supersede legacy structured data projects. A study by Levis et. al. shows improved clinical suicide risk prediction by including NLP-derived information from unstructured clinical notes.[35] Zhang et. al. use NLP-derived features from unstructured data to improve a structured data deep learning model.[25] A study from Massachusetts General Hospital shows improved fit in a suicide and accidental death after discharge model with NLP-derived features over structured features.[49]

Researchers also use NLP to prepare unstructured data for RFA and/or predictive modelling tasks. For example, the VUMC WordcloudIndexer, developed by Mandani et. al., uses a pattern-matching algorithm called regular expressions[50], to parse clinical notes, extract medical terms, map terms to Unified Medical Language System[51] (UMLS) concept unique identifiers (CUIs), and aggregate CUI counts from unstructured clinical notes.[52] The WordcloudIndexer provides timely, up-to-date access to structured, bag-of-words representations of clinical notes. Access to these data allows researchers to analyze unstructured clinical note information with traditional structured methods. In **Chapter 2** and **Chapter 3**, we use medical concepts from the WordcloudIndexer to lay the foundation for our analyses.

More advanced techniques instead tokenize unstructured texts into numeric arrays to preserve word order. Order-preserved representations are popular for used in word-embedding models[53,54] and feed-forward neural networks[55]. Word-embedding models, like word2vec[53], use ordered, tokenized representations of texts to learn vector-space embeddings for words in the corpus. Researchers use learned word-embeddings to transform texts into arrays of vectors which carry more relative meaning than an array of numbers. Vector-embedded texts are useful for analysis and feed-forward neural network training. Another embedding model, cui2vec, uses a combination of embedding techniques and transfer learning to create vector embeddings for UMLS[51] CUIs.[54] Deep neural networks (DNNs) trained with order-preserved text representations can yield high performance.[53,54,56] But, when model complexity exceeds training data quantity, overfitting becomes likely[57]; thus, DNNs (high complexity) are best suited for very large datasets.

Researchers can also use NLP to improve suicidal phenotype ascertainment, supporting label-dependent analyses like RFA or machine-learning-based risk prediction. For example, one study on a cohort of pregnant women shows an 11-fold increase in suicidal ideation ascertainment with NLP of clinical notes compared to diagnostic codes.[48] Other NLP applications include concept extraction,[58] sentiment analysis,[59,60] and topic assignment[61].

Network Analysis

In 1736, Leonhard Euler famously solved the Seven Bridges of Königsberg puzzle, proving that it was impossible circumnavigate the Prussian city of Königsberg in such a way to cross its seven bridges once and only once.[62] His proof[62] uses a diagram of dots and lines to represent the city – one of the earliest mathematical examples of a *graph*. Euler's work laid the foundation for graph theory, a subfield of mathematics which studies graphs.

Graphs are finite collections of vertices and edges.[63] Graphs can have any nonzero number of vertices (distinct points) and vertices can be connected by any number of directed or undirected edges.[63] Sociologists were among the first to use graph theory to model and study real world networks. In 1934, Moreno's work "Who Shall Survive?" used human-relationship-graphs, dubbed 'sociograms', to model relationships between children as they aged from kindergarten to 8th grade.[64] These early applications of graph theory were called social network analyses and eventually gave rise to network theory and network analysis, which study broader applications of graph theory beyond sociology.[65] Recent biomedical informatic uses of network analysis include analyzing surgical team collaborations,[66] healthcare interruptions,[67] and disease-gene relationships.[68]

The network analysis workflow usually consists of inference and analysis. *Network inference* often refers to using statistical evidence to infer regulatory networks between biological processes[69]; however, in this study we use 'inference' to refer to *any* inferential process used to generate a graphical network. Inference methods vary depending on the type of network being generated.[70] Biological networks often use gene expression reads to create coexpression edges between genes[69], while scientific literature networks instead use citations to create referential edges between publications.[71] Our research infers network edges between medical concepts (vertices) from intra-document term cooccurrence. We measure the frequency with which pairs of terms cooccur (occur within the same document) and generate edges for the most frequently cooccurring terms.

Network analyses commonly use network statistics from graph theory to glean insights from networks.[65,72] Local statistics, such as *degree* and *centrality*, characterize single vertices within a network, while global statistics, such as *density* and *diameter*, characterize the overall network with a single value.[70,72] Degree measures the number of 'incident' (i.e., connected) vertices to a specific vertex, and centrality measures how central a vertex is to the network.[70,72] Density measures the proportion of edges to possible edges within a network, and diameter measures the single longest path within a network.[70,72] These statistics are used to compare related networks to one-another, as well as the vertices within the same network. An added benefit of network statistics is that they give a multifactorial view of association. Degree and centrality both characterize vertex interconnection, but degree ranks general connectedness and centrality ranks significance within a cluster of connection.[70,72] Past researchers have applied network characterizations to derive valuable insights on a wide range of topics, including disease-gene relationships[68], posttraumatic stress disorder[73], and suicidality[36–40].

In our literature review, we found no studies on network analysis and suicide which used risk factors from clinical notes. The studies we reviewed used psychological assessments, questionnaires, and claims data to indicate risk factor presence. Although these data sources are clinically relevant and have high fidelity, they are difficult to collect and poorly serve at-risk individuals who have not undergone mental health treatment. Additionally, as noted earlier, suicide risk factors are poorly represented in structured EHR data (e.g., claims records)[47], but clinical notes have been shown to improve ascertainment of suicidal phenotypes.[48]

Supervised Learning

Early identification of suicidal individuals is a critical factor for successful intervention.[17,18] It may be possible to improve early identification of suicidal behavior with predictive modelling. In previous work at VUMC, researchers developed VSAIL, a supervised learning model (SLM) which was trained on structured EHR data to predict future suicide attempt and ideation.[23]

Supervised machine learning is an algorithmic predictive modelling technique which uses labelled data to train an algorithm to predict the probability of an outcome.[74] Perhaps the oldest and simplest SLM is simple linear regression, which generates a line of best fit for a given set of data points assumed to be linearly related.[75] The line of best fit is a straight line defined by two parameters: y-intercept and slope. These parameters are used to estimate the unknown value of the dependent variable y given the known value of the independent variable x.[75] Evolving from simple linear regression, multiple linear regression, allows for multiple independent variables and uses a number of parameters equal to the number of independent variables plus one.[76]

Logistic regression uses the same number of parameters as multiple linear regression, but uses the logistic function instead of the linear function to model the dependent variable.[77] L1 and L2 penalized regression models are logistic regression models with additional penalty terms to regularize the model weights.[78] Neural networks are layers of logistic regression models stacked together, sometimes with regularization. Each step in this progression represents an increase in both model complexity and model performance. The downside of complex models is that they require larger training datasets to avoid overfitting. Thus, care must be taken to select SLMs of appropriate complexity given available training data.

Research Questions & Approaches

Research Questions

(1) Can NLP-derived concept cooccurrences be used to generate suicide risk networks?
(2) If so, how can we improve upon prior attempts to create suicide risk networks?

(3) Can NLP-derived features from clinical notes improve an operational suicide risk prediction model?
(4) If so, what clinical terms are most important for assessing suicide risk?

Approach: Network Analysis

In **Chapter 2**, we explore the use of NLP-derived concept cooccurrences to generate suicide risk networks. We extract concept cooccurrences from clinical notes and use them to infer a risk factor networks. We use network statistics to analyze our networks and risk factors therein. We attempt to improve upon prior suicide risk network studies by using clinical-note derived features, building case-controls into our networks, and comparing multiple suicidal phenotypes.

Approach: Supervised Learning

In **Chapter 3**, we explore the use of NLP-derived concept counts to generate a suicide risk prediction model. Further, we train fusion models on combinations of clinical note-derived and structured data to test whether presently deployed models (VSAIL) can be improved. We compare multiple SLMs to optimize performance and select the top performing models for further analysis. To characterize which clinical terms are most important for assessing suicide risk, we quantify the feature importances of our top performing clinical note model.

Chapter 2: Exploring Risk Factors in Suicidal Ideation and Attempt Concept Cooccurrence Networks

KJ Krause, BS[1], John Shelley, BS[1], Alex Becker, BS[1], Colin Walsh, MD, MA[1]

[1]Vanderbilt University Medical Center, Nashville, TN, Department of Biomedical Informatics

Abstract

Suicide is a significant and rising threat to public health. In the United States, 47,500 people died from suicide in 2019, a 10-year increase of 30%. Many researchers are interested in studying the risk factors associated with suicidal ideation and suicide attempt to help inform clinical screening, intervention, and prevention efforts. Many suicide risk factor analyses draw from clinical subdomains and quantify risk factors independently. While traditional modeling approaches might assume independence between risk factors, current suicide research suggests that the development of suicidal intent is a complex, multifactorial process. Thus, it may be beneficial to how suicide risk-factors interact with one another. In this study, we used network analysis to generate visual suicidality risk relationship diagrams. We extract medical concepts from free-text clinical notes and generate cooccurrence-based risk networks for suicidal ideation and suicide attempt. In addition, we generate a network of risk factors for suicidal ideation which evolves into a suicide attempt. Our networks were able to replicate existing risk factor findings and provide additional insight into the degree to which risk factors behave as independent morbidities or as interacting comorbidities with other risk factors. These results highlight potential avenues for risk factor analyses of complex outcomes using network analysis.

Introduction

Suicide is a major cause of death and disability, both within the United States and abroad.[1] Globally, suicide accounts for 1.4% of all deaths, taking the lives of 800,000 each year.[2] In the United States, 47,500 people died from suicide in 2019.[3] Leading frameworks understand suicidality as a spectrum of passive ideation to active suicidal intent.[1,4–8] Suicidal ideation is more prevalent than suicidal attempt[9], and suicidal intent evolves from passive ideation through accumulation of practical capability (suicide means) and acquired capability (suicide desensitization).[5]

Many studies have sought to identify individual risk factors which mediate this process.[10–15] Mann et al. reported that the most effective interventions for preventing suicide attempt were physician education and lethal means restriction.[11] Spokas and colleagues reported that depression, hopelessness, childhood sexual abuse, and alcohol use disorder were all associated with impulsivity and likelihood of suicide attempt.[12] Other research reports that gender and sexual identity also affect the likelihood of suicide attempt.[13,14] Many of these analyses are limited by assumptions of independence.[15] The primary aims of this study are to expand the risk factor search to a broader range of medical histories and to analyze their cumulative impacts on suicide risk as a network rather than as isolated factors.

The electronic health record (EHR) is a rich source of patient information. Much of this information is contained within structured fields, such as diagnosis codes, medication lists, and problem lists.[16,17] Numerous studies have leveraged structured EHR data to advance prediction of suicide attempt and ideation.[16,18] However, emerging evidence shows that structured EHR fields lack acuity and coverage of data related to mental health.[19–22] This evidence also suggests that free-text clinical notes have greater coverage of mental health information and show promise for improving suicidality analyses.[19–22] The Vanderbilt Wordcloud Indexer uses regular expressions[23], a natural language processing (NLP) technique, to extract medical concepts from clinical notes.[24] Medical concepts are mapped to concept unique identifiers (CUIs) using the Unified Medical Language System (UMLS).[25]

Network analysis uses spatial representation of information to help visualize patterns and relationships within data.[26] In this study, we use network analysis to visualize how particular medical histories affect the risk-levels of others with respect to suicidality. Similar techniques have been used to analyze surgical team collaborations[27], healthcare interuptions[28], and disease-gene relationships[29]. In these examples, researchers sought a wider top-down view of the relationships within the systems they were studying. Similarly, we are interested in viewing how suicide risks operate together as a network.

Past researchers have used network analysis to generate suicide risk networks from psychological assessments and questionnaires[30–32] and to validate newly proposed psychiatric conditions[33]. We found no network analysis studies which used risk factors from clinical notes to study suicide. The studies we reviewed used psychological assessments, questionnaires, and claims data to indicate risk factor presence. Despite being clinically relevant and high fidelity, these data are difficult to collect and poorly serve at-risk individuals who have not undergone mental health treatment. Further, suicide risk factors are poorly represented in structured EHR data (e.g., claims records)[20], and clinical notes have been shown to improve ascertainment of suicidal phenotypes.[21] Thus, we attempt to improve upon earlier analyses by mining risk factors directly from clinical notes.

We use the Vanderbilt Wordcloud Indexer to extract frequencies of medical concept cooccurrences in the free-text health records of VUMC patients with coded histories of suicidal ideation and attempt. We generate risk network graphs, with nodes indicating individual medical concept associations with the outcome, and edges indicating pairwise cooccurrence associations with the outcome. We track three outcomes: suicidal ideation, suicidal attempt, and progression of suicidal ideation to attempt.

Methods

Data Sources

We mined structured patient data from the VUMC research derivative, EHR clinical data repository. We gathered medical concept occurrences from the Vanderbilt Wordcloud Indexer[24]. We downloaded the UMLS[25] data dictionary to create mappings between concept unique identifiers (CUIs) and their medical definitions. Approval for this study was obtained from the institutional review board (IRB).

Outcomes and Cohort

We tracked three outcomes: suicidal ideation, suicide attempt, and suicidal ideation progressing to suicide attempt. We extracted dates of coded suicidal ideation and suicide attempt from visit occurrences with International Classification of Disease (ICD) codes.[17] We used ICD-9 codes for visits occurring before October 2015, and ICD-10 codes for visits occurring after. We defined suicidal ideation progressing to suicide attempt as the occurrence of a suicidal ideation code prior to a suicide attempt code. The suicidal ideation and suicide attempt control cohorts were composed of the patients who did not have any coded record of suicidal ideation / suicide attempt, respectively. The progression control cohort was composed of all patients with coded record of suicidal ideation and no coded record of suicide attempt.

Data Preprocessing and Censoring

The wordcloud indexer had a vocabulary size of ~400k CUIs. The computational demands for calculating cooccurrences scaled exponentially to the vocabulary size (Suppl. A). Therefore, we used F-score correlation between each term in the vocabulary and each of our outcomes to create reduced vocabularies of the top-associated 1,000 CUIs with each outcome, dramatically reducing computational load (Suppl. A) while still analyzing a broad spectrum of possible risk factors. For each outcome dataset, we right-censored patient visits at the first recorded instance of each outcome to avoid contaminating our analysis of risk factors proceeding an outcome with concepts resulting from the outcome itself.

Differential Cooccurrence Matrix Generation

We built outcome-control differences directly into our network graphs to help visualize significant differentiating network associations without a separate control network graph. For each outcome and outcome-control, we had a dataset of per-document CUI counts with a vocabulary size of 1k, right-censored by outcome past the first recorded instance. The cooccurrence vocabulary size was then 9.99×10^5 (Suppl. A). We summed the total same-document cooccurrences for each CUI pair, and then summed the cooccurrences across every non-censored document for each patient. Next, we computed the count-per-million (CPM) normalized frequencies of the cooccurrence counts and scaled the frequencies to zero mean and unit variance. Finally, we calculated differential frequencies by subtracting the scaled and normalized CUI frequencies of each outcome's control from those of the outcome. (Suppl. B)

Network Analysis

The network generation steps below were repeated for each study outcome: suicidal ideation, suicide attempt, and suicidal ideation progressing to attempt First, we selected the top 25 CUIs by chi-squared correlation with the differential outcome (e.g., I_1 - I_0) to be starting nodes in the network. Next, we set an arbitrary edge-weight threshold (to be tuned later) to set inclusion criteria for edges and additional nodes in the network. We added cooccurrences with differential frequency above the inclusion threshold to the network as edges. Edge thickness was made proportional to magnitude of differential frequency. If any of the included cooccurrences included a CUI not selected in the original top 25, these CUIs were added to the network as additional nodes. We iteratively experimented with different edge-weight thresholds until the resulting networks appeared subjectively dense without excess clutter (approximately 25-30 nodes in the final network). Additionally, we added node colors to indicate individual CUI correlation with each given outcome.

We calculated the degree, degree-centrality, and eigen-centrality of each node in each outcome network. Degree measures the number of nodes a given node is connected to.[26] Degree-centrality and eigen-centrality are two different approaches for measuring the influence of nodes within a network; highly central nodes are nexuses of connection and influence within a network.[34,35] We also calculated the density and transitivity of each outcome network. Density measures the ratio of actual edges to possible edges in a network, and transitivity measures the overall extent to which a network is clustered.[26] We used Python 3.9 and the NetworkX package to conduct our network analysis.[36,37]

Results

In the suicide ideation cooccurrence network, the highest independently correlated terms were bipolar disorder, and post-traumatic stress disorder (PTSD), by chi-squared association. The highest degree terms were HIV, depressive disorder, and amphetamine abuse. The highest degree-centrality terms were the same. The top 10 terms by degree and degree-centrality included 7 psychiatric terms (depressive disorder, amphetamine abuse, anxiety, bipolar disorder, major depressive disorder, cocaine abuse, and alcoholism) and 3 non-psychiatric terms (HIV, colonoscopy, and dislocations). The network density was 0.10 and the network transitivity was 0.29. (Suppl. C, Figure 1)

Suicidal Ideation to Attempt Network Node Metrics

UMLS Concept	Degree	Degree Centrality	Eigen Centrality	Chi-Squared
Bipolar I Disorder	10.00	0.42	0.52	0.79
Electroconvulsive Therapy	8.00	0.33	0.40	6.52
Prediabetes Syndrome	7.00	0.29	0.40	1.15
Chronic Alcoholic Intoxication	6.00	0.25	0.03	0.96
Tobacco Use Disorder	4.00	0.17	0.01	0.61
Ataxia	4.00	0.17	0.01	3.41
Rib Fractures	3.00	0.12	0.00	9.68
Insomnia	3.00	0.12	0.27	0.04
Impulsive Behavior	3.00	0.12	0.27	0.22
Bipolar Disorder	3.00	0.12	0.27	0.10
Anxiety	3.00	0.12	0.27	0.07
Sleep Apnea, Obstructive	2.00	0.08	0.19	0.47
Amphetamine Abuse	2.00	0.08	0.19	0.47
Depressive Disorder	2.00	0.08	0.09	0.04
Acute Kidney Injury	2.00	0.08	0.19	0.08
Colonoscopy	1.00	0.04	0.00	0.28
Maxillary Fractures	1.00	0.04	0.00	6.95
Active Suicidal Ideation	1.00	0.04	0.08	0.01
Cont. Positive Airway Pressure	1.00	0.04	0.11	0.00
Contusion Of Lung	1.00	0.04	0.00	2.55

Table 1: This table gives the degree, degree-centrality, eigen-centrality, and chi-squared outcome correlation for the top 20 nodes in the ideation-to-attempt network, ranked by degree. The top 3 values for each metric are given in red text.

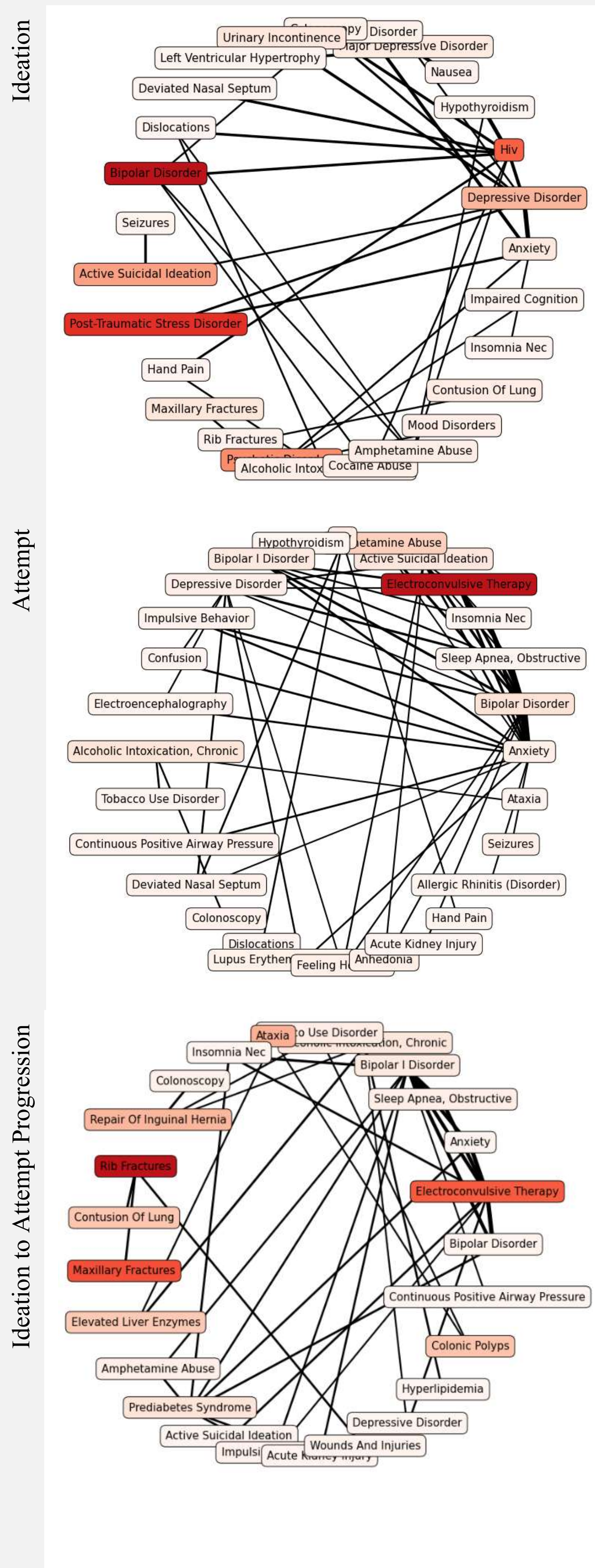
In the suicide attempt cooccurrence network, only electroconvulsive therapy (ECT) ranked highly in independent correlation with suicide attempt. The highest degree terms were anxiety, bipolar disorder, and depressive disorder. The highest degree-centrality terms were the same. The top 10 terms by degree and centrality included 8 psychiatric terms (anxiety, bipolar disorder, depressive disorder, ECT, active suicidal ideation, bipolar I disorder, insomnia, and feeling hopeless) and 2 non-psychiatric terms (obstructive sleep apnea and HIV). (Figure 1). The network density was 0.13 and the network transitivity was 0.24. (Suppl. D, Figure 1)

In the ideation-to-attempt progression network, the highest independently correlated terms were rib fractures, maxillary fractures, and ECT, by chi-squared association. The highest degree terms were bipolar I disorder, ECT, and prediabetes syndrome. The highest degree-centrality terms were the same. The top 10 terms by degree and degree-centrality included 7 psychiatric terms (bipolar I disorder, ECT, chronic alcohol intoxication, tobacco use disorder, insomnia, impulsive behavior, and bipolar disorder) and 3 non-psychiatric terms (prediabetes syndrome, ataxia, and rib fractures). The network density was 0.10 and the network transitivity was 0.29. (Table 1, Figure 1)

Discussion

We have developed an approach to identify clinical term cooccurrence networks for suicidal ideation and suicide attempt. For these primary outcomes, we computed the differences between scaled, normalized cooccurrence frequencies in the case (outcome) and control (no outcome) cohorts. We also identified clinically relevant ideation-to-attempt risk factors by computing the differences of scaled, normalized cooccurrence frequencies between the ideation-to-attempt cohort and the ideation-with-no-attempt cohort. The latter experiment parallels leading psychosocial ideation-to-attempt suicidality frameworks.[1,4–8]

Suicidal Ideation and Attempt Cooccurrence Networks



The highest degree terms in the ideation network were human immunodeficiency virus (HIV), depression, anxiety, amphetamine abuse, and bipolar disorder. HIV was the node with the highest degree overall, and is known to be a significant risk factor for suicide.[38] Patients with HIV have a 100-fold risk of suicide than the general population and in one study, only one general medical condition—traumatic brain injury (TBI)—was found to confer higher risk for suicide.[20,38]

The highest degree terms in the attempt network were anxiety, bipolar disorder, depression, ECT, and suicidal ideation. Interestingly, obstructive sleep apnea (OSA) was one of the most significant co-occurrences as well. Boggs et. al. found that sleep disorders (odds ratio [OR] for attempt = 3.1-4.3) were ranked third in conferring suicide risk only behind TBI (OR = 7.7-23.5) and HIV (OR = 1.4-6.0).[20] HIV dropped down on the degree rankings, although was still in the top 25. ECT, one of highest degree concepts and the concept with the highest single concept association, is used to treat treatment-resistant depression and has been shown to reduce suicidality.[39] The top-10 cooccurrence nodes in the attempt network included the term *feelings of hopelessness.* This finding is supported by "the Hopelessness Theory of Suicidal Ideation", a well-regarded psycho-sociological model for assessing suicidality.[40]

In the progression network, the highest degree terms were bipolar disorder, ECT, and prediabetes syndrome. The presence of ECT in the top degree rankings may be due to confounding as ECT is a common treatment for severe treatment resistant depression. In a meta-analysis of 27 studies comparing 12 demographic and clinical variables between suicidal ideation and suicide attempt, anxiety disorders, PTSD, drug use disorders, and sexual abuse history were the only variables significantly elevated in attempters compared to ideators.[1]

Figure 1: These graphs represent the associations of medical concepts and concept cooccurrences with the three study outcomes: suicidal ideation, suicide attempt, and suicidal ideation progressing to suicide attempt. Network nodes represent CUIs. Node color indicates chi-squared correlation of the given CUI with the given outcome. White is low association, red is high. Edges represent cooccurrences which are more frequent in patients in the case group than in the control group. Edge thickness is proportional to the difference between the cooccurrence frequencies in the case and control groups.

Similarly, the top 20 terms by degree ranking in the progression network include anxiety, drug use (alcohol, tobacco, and amphetamine abuse), and active suicidal ideation. Multiple studies have found associations between sleep disorders and suicide risk.[41,42] Interestingly, the progression network also included multiple terms associated with sleep disorders: insomnia, obstructive sleep apnea, and continuous positive airway pressure.

The centrality-ranked terms in our networks showed similarity to and overlapped with the individually-ranked terms – both replicated known associations with suicidal ideation and attempt. The differences between the two rankings suggest that some medical histories are more significant as comorbidities than morbidities and vice-versa. For example, in suicidal ideation, HIV is more significant as a comorbidity than bipolar, which is more significant as a morbidity. Further, in the progression network, we observed multiple indicators of sleep disorders, an observation consistent with existing research; however, our network also provides the additional observation that these sleep disorder terms are more significant as morbidities than as comorbidities. The ability to analyze risk factors in terms of morbidity and comorbidity highlight a major strength of network analysis.

We found no studies before ours that used risk factors from clinical notes to perform network analysis on suicide. The studies we reviewed used psychological assessments, questionnaires, and claims data instead to indicate the presence of risk factors.[30–33] Psychological assessments and questionnaires are clinically relevant and high fidelity, yet they can be difficult to collect. They also poorly serve at-risk individuals who have not gone through mental health treatment. Thus, our research fills an important gap in the research on suicide network analysis by using clinical notes to mine suicide risk factors, and producing a network model with scalable, readily available, high-quantity data such that it could be quickly and easily replicated in different settings without the need for manual data collection.

Further, we found only one study[31] before ours which used a control network in their network analysis of suicide. The control network helps deduce which network associations are attributable to suicidality specifically, rather than the study population. For example, there could be common mental health trends among veterans which appear in the suicidal phenotype network but would also appear in a control network of non-suicidal veterans. This is an example of selection bias, which could be mitigated with a control-network. Gijzen et. al. compared against a control network in their analysis of adolescents to help identify patterns specific to ideation, rather than to adolescence.[31] Our study fills a gap in the research by building case-control differences directly into our networks with edges determined by differential case-control associations.

Our study is limited by its use of observational data collected in real-world settings. Some likely sources of bias in our analysis include differences in healthcare utilization and frequency of screening for suicidal ideation amongst patient populations. The handful of resulting network terms (e.g., rib fractures) with little apparent connection to suicidality by research or intuition may be indicative of these differences. In addition, our outcomes are rare relative to the size of our cohort, potentially decreasing our power to detect true associations. Further, our study extracts outcome labels using both ICD-9 and ICD-10 codes for visits before and after October 2015, respectively, though previous validation studies have shown that ICD-10 is superior to ICD-9 for ascertainment of suicidal ideation and suicide attempt.[18]

Conclusion

In this study, we have developed a method that used network analysis to better understand free-text data located in the electronic health record. We showed that this method can re-capitulate known associations between clinical concepts and suicidality and can identify new ones. Future work is needed to quantify the ability of these concepts to predict progression from suicidal ideation to attempt.

Appendix

Supplement A: Computational Demands

Cooccurrence counts are computed by summing the total number of times each possible pair of CUIs occur together in the same document, for each patient. The computational demands are therefore proportional to the total number of possible CUI pairs to sum. The number of possible pairs N between two sets m and n is given by $N = m \times n$. In our case, n and m are the same vocabulary, hence $N = m \times m = m^2$. Finally, we subtract the number of self-cooccurrences, since they do not count as cooccurrences in our study, yielding $N = m^2 - m = m(m-1)$. The relationship between m and N is approximately given by $N \sim m^2$. By reducing the vocabulary size m from 400,000 to 1,000, we reduce the number of possible cooccurrence pairs N from 1.6×10^{11} to 1×10^6, a 1.6×10^5 fold reduction in computational demand.

Supplement B

Differential-Frequency Cooccurrence Matrices Equation

$$\Delta\sigma(CPM^{m,n}) = \sigma(CPM^{m,n}_{Y=1}) - \sigma(CPM^{m,n}_{Y=0})$$

m: term 1
n: term 2
Y: outcome (1 = case, 0 = control)
Ϭ: Scaling operation (zero mean, unit variance)

This equation shows the mathematic procedure used to obtain differential frequencies for each outcome. This procedure is repeated for each outcome: suicidal ideation (I), suicide attempt (A), and suicidal ideation progressing to suicide attempt (P).

Supplement C

Suicidal Ideation Network Node Metrics

UMLS Concept	Degree	Degree Centrality	Eigen Centrality	Chi-Squared
HIV	9.00	0.35	0.52	32.70
Depressive Disorder	7.00	0.27	0.27	16.72
Amphetamine Abuse	6.00	0.23	0.41	2.66
Anxiety	5.00	0.19	0.10	4.14
Bipolar Disorder	4.00	0.15	0.34	49.69
Major Depression	4.00	0.15	0.09	3.68
Colonoscopy	3.00	0.12	0.26	1.22
Chronic Alcoholism	3.00	0.12	0.19	2.29
Cocaine Abuse	3.00	0.12	0.29	1.66
Psychotic Disorders	3.00	0.12	0.02	24.40
Dislocations	3.00	0.12	0.25	0.04
Active Suicidal Ideation	2.00	0.08	0.07	20.53
PTSD	2.00	0.08	0.08	41.18
Hand Pain	2.00	0.08	0.16	0.56
Urinary Incontinence	2.00	0.08	0.08	4.66
Hypothyroidism	2.00	0.08	0.21	0.15
Rib Fractures	2.00	0.08	0.00	1.08
Contusion Of Lung	1.00	0.04	0.00	2.98
Mood Disorders	1.00	0.04	0.01	2.42
Nausea	1.00	0.04	0.02	1.06

This table gives the degree, degree-centrality, eigen-centrality, and chi-squared outcome correlation for the top 20 nodes in the ideation network, ranked by degree. The top 3 values for each metric are given in red text.

Supplement D

Suicide Attempt Network Node Metrics

UMLS Concept	Degree	Degree Centrality	Eigen Centrality	Chi-Squared
Anxiety	15.00	0.56	0.42	4.21
Bipolar Disorder	12.00	0.44	0.42	15.36
Depressive Disorder	9.00	0.33	0.30	5.59
Electroconvulsive Therapy	7.00	0.26	0.34	114.05
Active Suicidal Ideation	7.00	0.26	0.31	10.96
Sleep Apnea, Obstructive	6.00	0.22	0.31	0.47
Bipolar I Disorder	5.00	0.19	0.26	9.46
HIV	4.00	0.15	0.02	7.07
Insomnia	4.00	0.15	0.22	0.40
Feeling Hopeless	3.00	0.11	0.16	3.63
Chronic Alcoholism	3.00	0.11	0.00	13.18
Deviated Nasal Septum	3.00	0.11	0.11	2.54
Amphetamine Abuse	2.00	0.07	0.13	25.26
Anhedonia	2.00	0.07	0.11	3.92
Lupus Erythematosus	2.00	0.07	0.11	3.06
Electroencephalography	2.00	0.07	0.11	1.78
Confusion	2.00	0.07	0.11	0.77
Impulsive Behavior	2.00	0.07	0.13	1.47
Hand Pain	1.00	0.04	0.00	0.18
Colonoscopy	1.00	0.04	0.00	2.26

This table gives the degree, degree-centrality, eigen-centrality, and chi-squared outcome correlation for the top 20 nodes in the attempt network, ranked by degree. The top 3 values for each metric are given in red text.

Chapter 3: Improving a Clinically Operational Suicide Risk Prediction Model with Natural Language Processing and Machine Learning

Kevin J Krause[1], Zhijun Yin[1], Sharon E Davis[1], Colin G Walsh[1]

[1]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN

Abstract

In 2019, approximately 47,500 people in the United States died of suicide. Interventions such as firearm access restriction and psychiatric medication prescription reduce suicide risk. However, these prevention efforts are only effective when at-risk individuals can be located. Vanderbilt University Medical Center currently identifies at-risk patients with the Vanderbilt Suicide Attempt and Ideation Likelihood risk model (VSAIL), an operational suicide prevention tool using structured electronic health record data to predict suicide attempts. This model does not consider free-text clinical notes, which have proven effective in many other clinical prediction models. To further improve model performance, we propose using existing natural language processing tools to integrate information from free-text clinical notes into the model. We ascertained suicide attempt events with billing codes and extracted bag-of-words representations of clinical notes from the Health Data Repository Initiative at VUMC. We trained machine learning models (logistic regression, random forest, and gradient boosting machine) to predict suicide attempts within 30 days of a hospital visit using the clinical terms present clinical notes from the prior 90 days. We trained an early-fusion model on the combined features from the structured and clinical note models, and a late-fusion model on the predictions made by the structured and clinical note models. Last, we compared the model performances (structured, clinical note, early-fusion, and late-fusion) on a common validation test set to select the optimal model. Our experiments on a 240k patient validation cohort demonstrated that the clinical note model outperformed the previously-developed structured model (average precision = 0.33 and 0.19, respectively; $p < 0.001$), and that the late-fusion model outperformed all other models (average precision = 0.41, $p < 0.001$). These results suggest that clinical notes alone contain rich information absent from the structured record and that the structured and clinical note data complement one-another as inputs to clinical prediction models.

Introduction

In 2019, approximately 47,500 people in the United States died of suicide.[1] For at risk individuals, there are proven interventions such as firearm access restriction and psychiatric medication prescription.[17] However, there are significant barriers to identifying at-risk individuals and delivering timely prevention efforts. Natural language processing (NLP) of clinical narratives captured during medical encounters might reduce this gap by identifying at-risk individuals, facilitating interventions, improving ascertainment of suicidality, and uncovering attempt trends.

The use of machine learning models in clinical decision support tools is growing in popularity amongst researchers at academic medical institutions.[89–91] This rapid increase in popularity is largely due to a growing body of evidence that machine learning models can significantly enhance disease prediction efforts across multitudes of clinical domains. Most clinical machine learning models are trained exclusively on structured data within the electronic health record (EHR). Structured data are highly organized, making them easier for machine learning algorithms to process than unstructured data, such as free-text clinical notes. However, structured fields within the EHR lack information present within the unstructured portions of clinical notes. Researchers have developed a myriad of useful NLP methods for extracting and structuring free-text information.[50,53,54,59–61] Many studies have now demonstrated the use of these NLP methods to improve the performance of structured data based clinical machine learning models.[25,35]

In previous work, researchers at Vanderbilt University Medical Center (VUMC) used structured data from the EHR to develop the Vanderbilt Suicide Attempt and Ideation Likelihood (VSAIL) model for predicting suicide attempt and ideation.[23,41] VSAIL was deployed in 2019 to VUMC's Epic-based EHR and generates suicide risk scores at the start of every patient encounter. VSAIL inputs are counts of occurrences for a range of diagnosis, medication, and problem codes within a patient's medical history. The current iteration of VSAIL does not include free-text clinical note data.[41] However, emerging research demonstrates a gap in the structured ascertainment of pertinent mental health markers, particularly around suicidality.[35,47–49] A study by Boggs and colleagues found that the lack of structured data on suicidality within the electronic health record (EHR) led to significant gaps in follow up assessments pertaining to suicidal ideation.[47] A study with the Veterans Health Administration used sentiment analysis to improve the prediction accuracy of their deployed suicidality prediction model.[35] A team of researchers employed NLP to ascertain suicidal ideation in pregnant women and noted an 11-fold increase in capture over diagnostic codes.[48] Investigators from Massachusetts General Hospital applied NLP to predict risk of suicide and accidental death after discharge and improved their model fit over a structured data model.[49]

We hypothesize that clinical note information will enhance clinically meaningful aspects of predictive model performance—both alone and in combination with the structured model. We consider multiple ways in which the structured and unstructured features may be incorporated. We use a cost-effectiveness perspective to compare models and provide practical insights into how different models may or may not be operationally impactful.[92]

Temporal Study Design

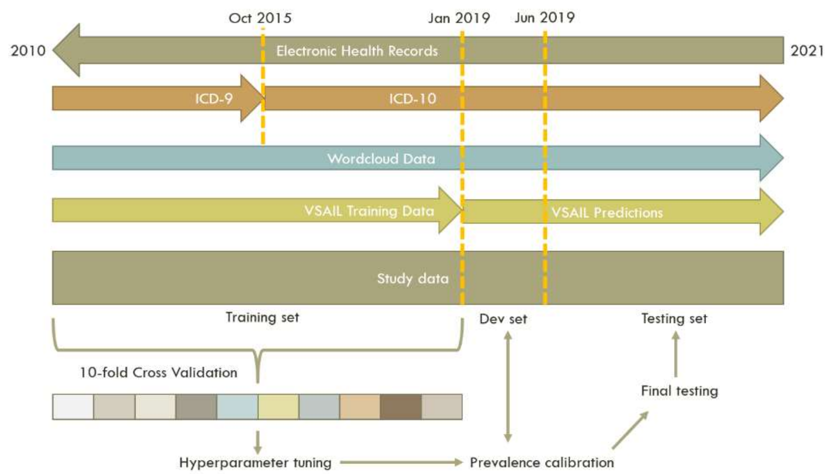


Figure-1: This figure depicts the temporal layout of our study design. EHRs are included from 2010 onward. The 2015 shift from ICD-9 to ICD-10 is indicated by an orange arrow occurring within the training set. Wordcloud data are included for the entire timeline. VSAIL data shifts from training to production in January 2019. The data are split into training, development, and testing sets according to the dates shown in the figure. The training set undergoes 10-fold cross-validation and hyperparameter tuning before prevalence calibration on the development set. Finally, the selected models are tested against the test set.

Methods

Cohort and Study Design

Figure-1 depicts our data collection window and the temporal layout of model training and validation. We collect data on all adult outpatient and emergency department (ED) encounters at VUMC between January 1, 2010 and December 31, 2021. Inpatient visits are excluded to stay consistent with the original development of the VSAIL model.[23] Encounters occurring within 3 days of one-another are grouped and treated as a single encounter. Clinical note data are available across the entire timeline, however production VSAIL predictions are only available from January 2019 onward. We use testing data only from after VSAIL's deployment date to avoid overlap between VSAIL training data and our final validation cohort. Additionally, on October 1, 2015, VUMC switched from using ICD-9 billing codes to using ICD-10 billing codes. This shift was significant, as ICD-10 demonstrated superior Positive Predictive Value (PPV) for identifying positive cases of suicide attempt.[24] To maximize the ground truth PPV for final validation, we sample testing data from after the shift to ICD-10 was completed. Thus, we temporally split our training, development, and testing datasets on January 1, 2010 to December 31, 2018, January 1, 2019 to August 31, 2019, and June 1, 2019 to December 31, 2021, respectively.

We define positive or negative indication of suicide attempt for each encounter by the presence or absence of an ICD billing code for suicide attempt within 30-days following the encounter. In the original development of VSAIL, a 30-day prediction window was determined by clinical experts to be optimal.[23,41] Encounters occurring within 3 days of a suicide attempt are dropped from analysis to avoid analyzing visits initiated by a suicide attempt. Remaining encounters with a valid suicide attempt indicator are right-censored 3 days prior to the attempt to avoid capturing data related to the attempt visit in the predictive model. The training set contains 155k encounters and 3.9k attempt observations. The validation set contains 241k encounters and 3.4k attempt observations.

Clinical Note Model

We extract medical concept counts from clinical notes using the VUMC Wordcloud Indexer, a regular expressions-based NLP tool developed by Giuse and colleagues.[52] The Wordcloud Indexer supports a vocabulary of ~400k medical concepts which are identified by concept unique identifiers (CUIs) in the Unified Medical Language System. The Wordcloud Indexer uses a bag-of-words representation, where syntactical structure from the clinical note is not preserved. The VUMC Wordcloud Indexer CUI extraction process is automatic and timely, such that these data would be available for any future real-time implementation of a clinical note model. We limit CUI collection to clinical notes within a 90-day time window preceding the prediction encounter.

Within the clinical note model preprocessing pipeline, we preform term total frequency inverse document frequency (TFIDF) transformation, k-best feature selection, majority undersampling, and minority oversampling. TFIDF is a scaling procedure used to weight terms inversely proportional to their total occurrences in the entire corpus to emphasize terms which occur less frequently.[93] We implement K-best feature selection[94] to reduce the number of input features (CUIs) to the model by ranking them according to F-score[95] correlation with suicide attempt and selecting the top ranked terms. We use majority undersampling and minority oversampling[96] to create a more balanced distribution of attempt and non-attempt observations in the training data. Both random and near-miss[97] undersampling are tested in the pipeline. The undersampling rate is also tested as a hyperparameter. The synthetic minority oversampling technique (SMOTE) is applied to generate synthetic observations of suicide attempt positive encounters, using a K-neighbors algorithm.[98] We tuned the oversampling rate and k-neighbors constant as hyperparameters.

We test several machine learning algorithms including logistic regression (LR)[77,99], L1-penalized LR[78], L2-penalized LR[78], gradient boosting machine[100], and random forest[101]. These algorithms are chosen for their interpretability, simplicity, and widespread acceptance within the clinical informatics domain.[77,101–104] We apply a randomized grid search to test combinations of preprocessing elements, preprocessing hyperparameters, and final classifiers with 10-fold stratified cross-validation.

Study Cohort Demographics (N = 106,338)			
Race	Total	Attempts	Percent
Black (15.0 %)	15,922	252	1.58 %
Asian (1.85 %)	1,967	22	1.12 %
White (79.5 %)	84,590	984	1.16 %
American Indian / Alaskan (0.28 %)	294	6	2.04 %
Pacific Islander (0.01 %)	11	1	9.09 %
Other (2.43 %)	2,582	15	0.58 %
Ethnicity	Total	Attempts	Percent
Non-Hispanic (94.4 %)	100,422	1,194	1.19 %
Hispanic / Latino (3.56 %)	3,786	52	1.37 %
Unknown (1.93 %)	2,050	14	0.68 %
Gender	Total	Attempts	Percent
Male (44.0 %)	46,771	558	1.19 %
Female (55.9 %)	59,483	702	1.18 %

Table-1: This table summarizes the overall coded demographics of the study, including suicide rates within each coded demographic. The percentage of the study population in each category is shown in parenthesis beside each category.

AUROC / AUPRC Comparison Against Validation Set

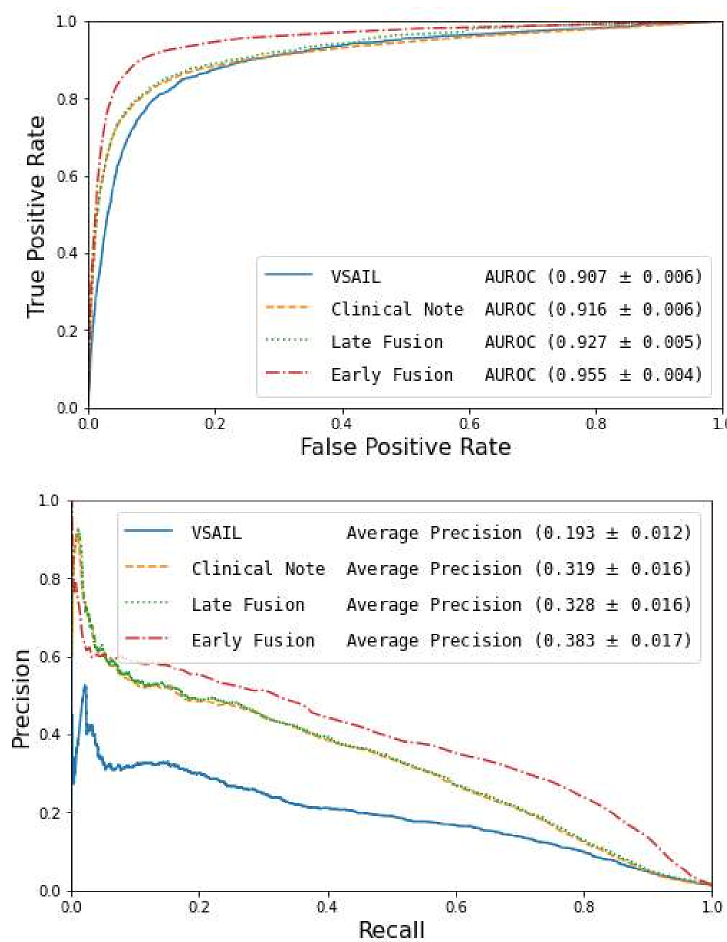


Figure-2: This figure compares the validation AUROC (top) and AUPRC (bottom) of each model.

The optimal model is selected by lowest average f1-score and lowest standard deviation across the cross-validation folds. We calibrate the final model to the prevalence of suicide attempt within the development set using Platt's method.[105] We measure model calibration before and after adjustment using Spiegelhalter's z-score.[106]

After model training and selection, we test our highest performing model on two patient cohorts: low and high healthcare utilization (1 visit in the past 90 days, > 1 visit in the past 90 days), to assess whether the 90-day window worsens performance within either group. The high and low threshold was set to match the median utilization within the test set.

Fusion Models

To test the impact of combining structured and clinical note data on modelling performance, we generate two fusion models: early-fusion and late-fusion. The early-fusion model is trained on the combined set of structured and clinical note model input features, using the same preprocessing pipeline and tuning strategy described above in *Clinical Note Model*. The TFIDF preprocessing step is only applied to the CUI counts and is not applied to the VSAIL features. The late-fusion model is generated by stacking the structured (VSAIL) and clinical note models through a single logistic regression classifier. Both ensemble models are calibrated as described earlier in *Clinical Note Model*.

Evaluation

We compare the performances of the four models described above: VSAIL, clinical note, early-fusion, and late-fusion. We assess general performance on the final validation set with Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision Recall Curve (AUPRC), also known as average precision (AP). The evaluation metrics are tailored to developing a suicide screening tool. AUPRC is prioritized over AUROC for its relevance with rare outcomes, such as suicide attempt. We measure model cost-effectiveness to evaluate potential clinical

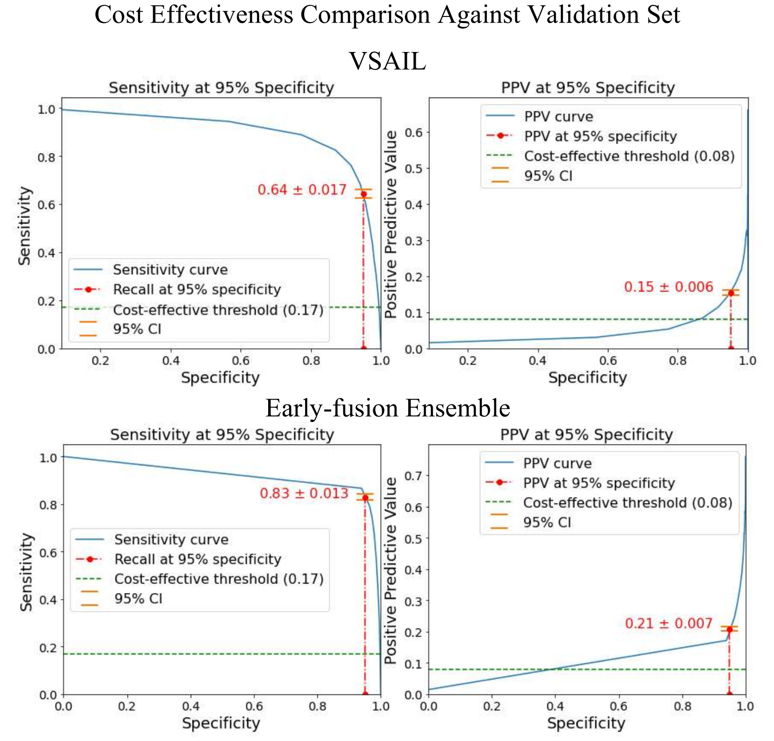


Figure-3: This figure compares the cost effectiveness curves for VSAIL and the early-fusion ensemble models. Cost effectiveness is measured in terms of sensitivity and PPV at 95% specificity, as described by Ross et. al., 2021. Minimum performance thresholds are indicated in green. The 95%-point estimate is shown in red, with a 95% confidence interval in orange, determined by bootstrapping.

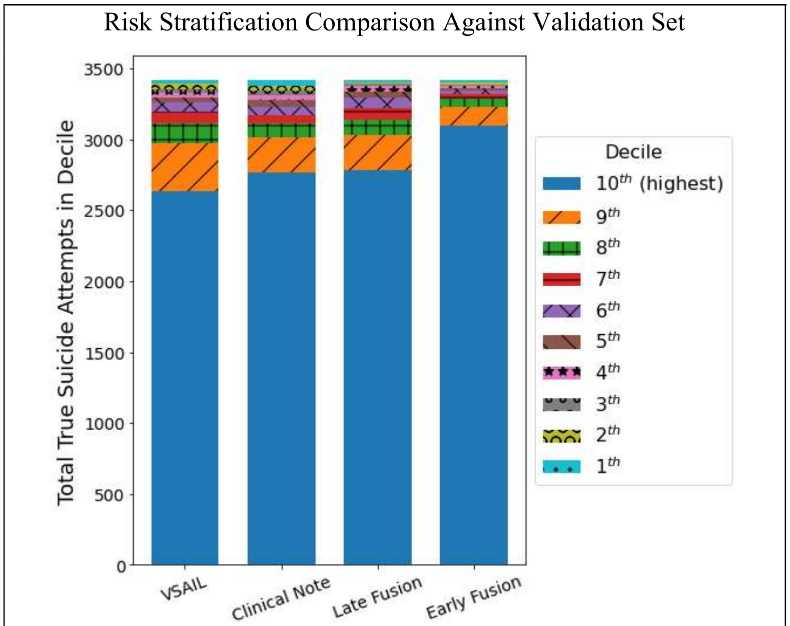


Figure-4: This figure depicts the number of suicide attempts binned in each probability decile of each model.

burden[92] by comparing sensitivity and PPV at 95% specificity, as described by Ross et al.[107]

We use bootstrapping with 1k iterations to generate 95% confidence intervals for the average AUROC, average AUPRC, sensitivity at 95% specificity, and PPV at 95% specificity. We evaluate risk stratification by counting true positives within probability deciles. The risk stratification plots provide a patient-centered view of how an improved model can translate value directly to the identification of additional future suicide attempts. To better understand the impact of clinical notes on our suicide prediction model, we use average mean decrease in impurity[108] across 100 bootstrapped variations of the clinical note model to calculate feature importances.

Results

Cohort Characteristics & General Performance

Table-1 describes the demographics of the entire study cohort. Figure-2 depicts the AUROC and AUPR curves, respectively, for each of the four models. Both AUROC and AUPR are highest for the early-fusion model ($p < 0.001$), followed by the late-fusion model and the clinical note model, with the lowest values observed for the VSAIL model.

Cost Effectiveness Analysis

Figure-3 depicts the cost effectiveness sensitivity and PPV curves at 95% specificity for the lowest and highest performing models, VSAIL and late-fusion ensemble, respectively. Ross et. al. conclude that suicide screening models should meet or exceed a sensitivity of 0.17 and PPV of 0.08 at 95% specificity. All four models pass this threshold by a wide margin. Sensitivity and PPV at 95% specificity are highest in the early-fusion model, followed by the late-fusion and clinical note models, with the lowest values observed for VSAIL.

Risk Stratification

Figure-4 depicts the risk stratification of true positive suicide attempts counts within outputted probability deciles. The early-fusion model assigns the highest number of suicide attempts within the highest probability decile (3,093 out of 3,879 suicide attempts). The other three models rank late-fusion (2,785), clinical note (2,767), and VSAIL (2,638). The total difference between the early-fusion ensemble and VSAIL is 445 additional suicide attempts within the highest probability decile over an 18-month period.

Figure-5 depicts the top-20 feature importances of the top performing model (early-fusion) by average mean decrease in impurity. We test the early-fusion ensemble model on high (upper 50th quantile) and low (lower 50th quantile) healthcare utilization portions of the test data and find higher AUROC in the high-utilization group ($p < 0.001$) and higher AUPRC in the low-utilization group ($p < 0.001$) (figure-6).

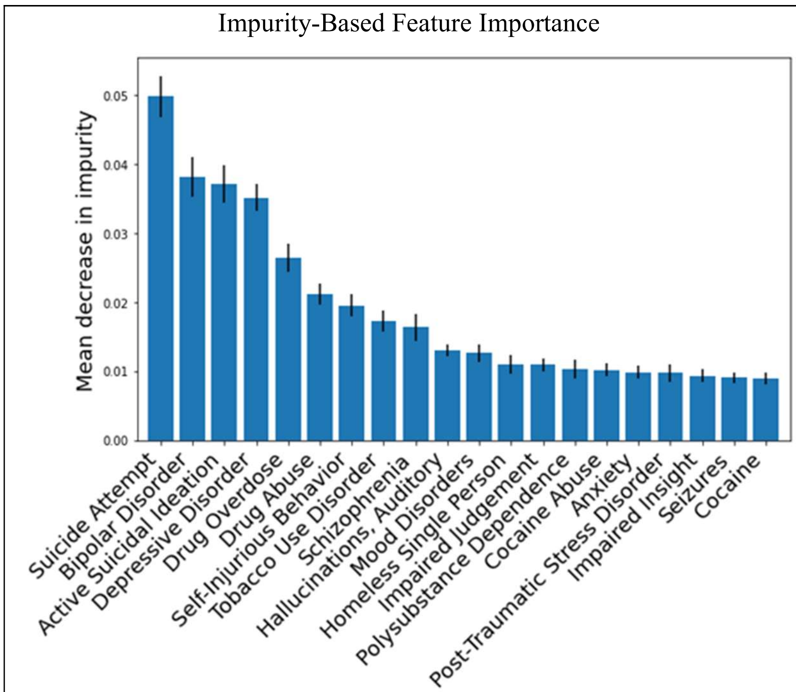


Figure-5: This figure depicts the top-20 most important features by mean decrease in impurity from the early-fusion model. Error bars illustrate the variance of impurity decrease across all the trees in the random forest.

Discussion

Across all testing dimensions (AUROC, AUPRC, cost-effectiveness, risk stratification) the early-fusion ensemble model outperforms all other models. Additionally, by most measures (AUPRC, cost-effectiveness, risk-stratification), the clinical note model outperforms VSAIL. The risk stratification test shows more than 400 additional suicide attempts captured within the highest risk probability decile between the early-fusion ensemble and

VSAIL over an 18-month period. In total, the inclusion of clinical note information into our suicide risk prediction model results in improved performance while reducing number needed to screen and improving cost effectiveness. The top importance features by mean decrease in impurity (MDI) include terms related to suicidality, mental health, depression, and drug use. In previous studies, many of these terms are found to be associated with increased suicide risk.[33] These results show that meaningful suicide risk indicators which are absent from structured fields in the EHR exist in clinical notes, and that we can successfully incorporate these predictive signals into an improved suicide risk prediction model.

A 30-day prediction window is chosen for this study based on clinical input; however, we do not test alternate time windows for their potential impacts on performance. It is possible that, despite clinical preference for a 30-day model, a different window setting would improve the model's performance or result in a different model (e.g., clinical note model) outperforming the others. Similarly, we do not test input time windows other than 90 days. We do show that performance varies slightly between high and low utilization groups, with higher AUROC in the high utilization group and higher AUPRC in the low utilization group. The difference between AUROC and AUPRC may be because the low prevalence of suicidality increases the true negative rate.

In the future, we plan to apply more advanced NLP methods to further improve our risk prediction model. For example, we believe that a vector embedding step, such as word2vec[53] or cui2vec[54], may prove beneficial. These vector embedding methods use cooccurrence-based context to transform terms into vectors which carry added information, such as similarity to other terms in the corpus. In addition, we would like to take steps toward clinical deployment of our model. Successful clinical deployment of a machine learning model is a highly challenging endeavor, which requires careful stakeholder collaboration, application development, validation, and maintenance.

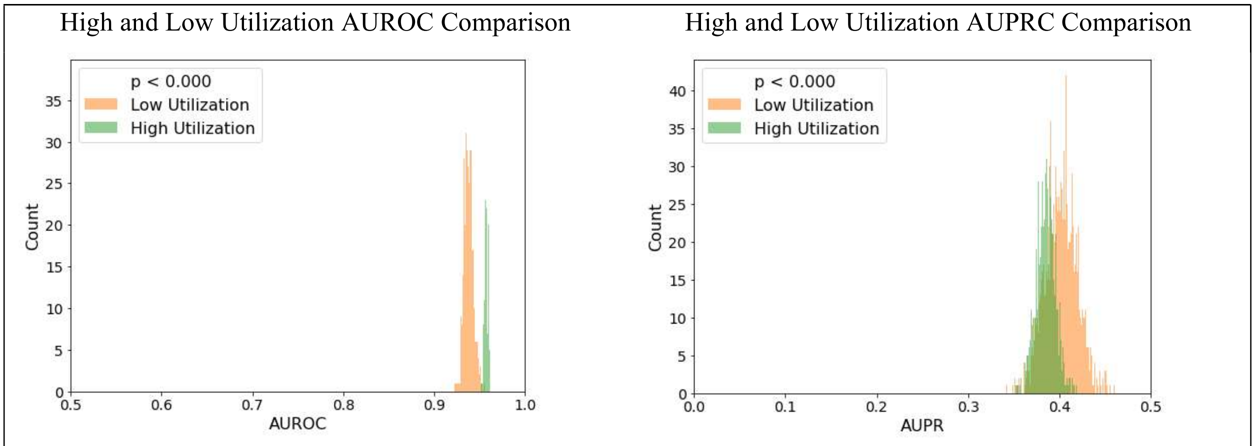


Figure-6: These two plots show the average AUROCs (left) and AUPRCs (right) between the low and high healthcare utilization groups (below and above 50th percentile, respectively) across 1,000 bootstrap iterations. The legend in each plot displays the p-value from a two-tailed t-test assuming unequal variance.

Chapter 4: Summary

Primary Findings

Network Analysis

Our network analysis identified several high-degree terms presumed to be comorbid risk factors for suicidal phenotypes: HIV, depression, anxiety, amphetamine abuse, obstructive sleep apnea, bipolar disorder, ECT, suicidal ideation, and prediabetes syndrome. Many of these associations, such as HIV and OSA, are supported by previous studies.[32,47,87,88] Further, our network analysis of suicidal ideators who progressed to suicide attempt replicates results from previous studies which show increased risk for suicide attempt relative to ideation: anxiety, drug use, active suicidal ideation, and sleep disorder.[13,87,88]

The centrality-ranked terms in our networks are similar to the degree-ranked terms; however, differences suggest that some medical histories are more significant as comorbidities than morbidities and vice-versa. For example, in suicidal ideation, HIV is more significant as a comorbidity than bipolar, which is more significant as a morbidity. Furthermore, the progression network reveals multiple indicators of sleep disorders, an observation consistent with existing research. In addition, our network also provides the observation that these sleep disorder terms are more significant as morbidities than as comorbidities.

Supervised Learning

Our early-fusion ensemble model outperformed all other models in our analysis. Additionally, the clinical note model outperformed VSAIL by most measures. The risk stratification test shows more than 400 additional suicide attempts captured within the highest risk probability decile between the early-fusion ensemble and VSAIL over an 18-month period. Using clinical note information in our model improved performance while reducing number needed to screen and improving cost effectiveness. The top importance features by mean decrease in impurity include terms related to suicidality, mental health, depression, and drug use; many of which are found to be associated with increased suicide risk in previous studies.

The results of these studies demonstrate that clinical notes contain suicide risk signals which can be extracted through concept counts. Many of these signals are absent from structured EHR fields. With these extracted clinical note signals, we can generate logical, intuitive suicide risk networks and train machine-learning models to predict the risk of future suicidal behavior.

Strengths and Contributions

Predictive Modelling

Prior studies have demonstrated reliable methods for ascertaining suicidal behavior from the EHR.[24,25] Other studies have built upon suicide ascertainment and have trained predictive models for suicidal behavior.[23] Many operational suicide prediction models have been trained on structured data[23]; however, studies[25,35,48] have shown that unstructured data can be used to improve predictive models. Our early-fusion model (structured/unstructured hybrid) and clinical note model (unstructured) both outperform VSAIL (structured). These results support earlier findings by other researchers that (1) unstructured training data can improve machine-learning models and (2) unstructured data are better at capturing suicidal behavior than structured data. Prior studies[35,49] have used sentiment analysis on unstructured clinical notes to improve structured-data-based suicide prediction models. These studies used sentiment analysis to represent clinical notes as multi-axis sentiment measurements, while we used pattern-based[50] extraction to represent clinical notes as unordered bags of CUIs[51]. By using a different methodology, our study (3) supports the hypothesis that model improvement is due to the presence of better suicidal behavior signals embedded in unstructured clinical notes.

Epidemiology

Past epidemiological studies[13,32,33,87,88] have used association measures and regressions to identify suicide risk factors. Epidemiological methods commonly rely on the assumption that their variables are independent[109] (i.e., not linearly correlated) and classify relative risk for each variable in isolation. Epidemiology studies use various statistical methods to address this limitation, such as by using analysis of variance (ANOVA)[110] to characterize the relationships between the dependent and independent variables.[109] In our study, we used independent association measures to perform a first-pass filter against our risk factors, before using term cooccurrence to characterize risk interdependence. Further, we used risk co-associations to pull new terms into the risk network graph. Our risk networks can differentiate risk by independent-association, degree, and centrality. Independent association gives a straightforward relative risk perspective, while degree and centrality confer comorbidity and risk-interdependence. Despite these differences, the top degree and centrality terms in our suicide risk networks have been found by epidemiological studies[13,32,33,87,88] to confer high relative risk for suicidal behavior. The big picture results from these analyses are therefore in general agreement. Our use of network analysis (4) supports the results of existing risk factor analysis while adding broader themes of risk interdependence to deepen our understanding of suicide risk.

Network Analysis

Our project is not the first to apply network analysis to the subject of suicide. Earlier we noted the significant contributions by de Beurs[36], Simons et. al.[37], Graziano et. al.[38], Gijzen et. al.[39], and Bloch-Elkouby et. al.[40], who pioneered the use of network analysis to study suicide. Among their contributions are the use of psychological assessments and questionnaires to generate suicide risk networks[37–39] and the use of network analysis to validate the newly proposed Suicide Crisis Syndrome[40]. Each of these studies made significant contributions to the relatively unexplored arena of analyzing suicidality with network diagrams. Still, their works leave a few interesting gaps.

First, we found no studies that used risk factors from clinical notes to do a network analysis of suicide. The studies we reviewed used psychological assessments, questionnaires, and claims data instead to indicate the presence of risk factors. Even though psychological assessments and questionnaires are desirable data sources because of how relevant they are clinically and how high fidelity they are, they can be difficult to collect. They also don't serve at-risk individuals well who have not gone through mental health treatment. Earlier we noted that suicide risk factors are not represented well in structured EHR data[47], and that clinical notes have been shown to improve ascertainment of suicidal phenotypes.[48] Thus, our research fills an important gap in the research on suicide network analysis by (5) using clinical notes to mine suicide risk factors, and (6) producing a network model with scalable, readily available, high-quantity data such that it could be quickly and easily replicated in different settings without the need for manual data collection.

Further, we found only one study[39] which used a control network in their network analysis of suicide. The control-network is an important facet which should be used to strengthen these analyses. Without a control to compare against, which associations are attributable to suicidality specifically, rather than the study population. For example, there could be common mental health trends among veterans which appear in the suicidal phenotype network but would also appear in a control network of non-suicidal veterans. This is an example of selection bias, which could be mitigated with a control-network. Gijzen et. al. compared against a control network in their analysis of adolescents to help identify patterns specific to ideation, rather than to adolescence.[39] Our study fills a gap in the research by (7) building case-control differences directly into our networks with edges determined by differential case-control associations. In addition, we found no studies which compared different phenotypes in their network analysis of suicide. Our study fills another important gap by (8) using a three-phenotype comparison to distinguish risk patterns between different suicidal phenotypes (i.e., attempt, ideation, and ideation progressing to attempt). These added factors help provide discernment to the suicidal behavior networks we are studying.

Contribution Summary List

(1) Supports earlier findings that unstructured training data can improve machine-learning models.
(2) Supports earlier findings that unstructured data are better at capturing suicidal behavior than structured data.
(3) Supports the hypothesis that model improvement is due to clinical notes.
(4) Deepens our understanding of suicide risk by visualizing risk interdependence.
(5) Uses clinical notes to mine suicide risk factors for network analysis.
(6) Produces a network model with scalable, readily available, high-quantity data.
(7) Builds case-control differences directly into our networks.
(8) Uses a three-phenotype comparison to distinguish risk patterns between different suicidal phenotypes.

Limitations

Selection Bias

The current study has several limitations that should be considered when interpreting its findings. First, we will discuss possible sources of selection bias, or bias resulting from study participation factors.[111] To remain consistent with the original development of VSAIL, we only included ED and inpatient encounters in our analysis. This decision allowed us to perform a 1:1 validation of our models against the original VSAIL model[23] developed by Walsh et. al. However, the exclusion of outpatient encounters could skew our data. Patients with emergency encounters may exhibit higher rates of suicidal behavior and different patterns of psychiatric distress than those without. Thus, our model could struggle to correctly classify patients with less severe suicidal ideation who have not made emergency visits. Another source of selection bias in this study is the exclusive use of data from VUMC. Vanderbilt, though situated in Nashville Tennessee, sees many patients from across the state who seek specialty treatments unavailable close to home. Further, Nashville and VUMC both have demographic makeups which differ from other regions and medical centers. Both factors could introduce selection bias into our study and make it less generalizable to other areas and institutions.

Information Bias

Next, we will discuss possible sources of information bias, or bias resulting from measuring health data.[111] First, this study trains models with ICD-9 codes, which have been shown to exhibit poor sensitivity and positive predictive value for ascertaining suicide attempt and ideation.[112] To address this, we validated our models exclusively with ICD-10 codes, which have been shown to exhibit reasonably high positive predictive values with suicide attempt and ideation ascertainment.[24] Despite this, ICD-10 codes are still imperfect indicators for suicidal behavior.[24] Second, our use of administrative claims data as structured model inputs could introduce misclassification bias.[113] For example, providers may code for psychiatric symptoms (e.g., anxiety, depression, anhedonia) differently depending on whether the patient is seeking pharmaceutical management and who their insurance provider is. Third, our structured data and clinical notes may underrepresent psychiatric conditions due to perceived stigma.[112] Fourth, physician errors and patient-reported histories could cause inaccuracies in clinical notes. Finally, differences in healthcare utilization and mental health screening may result in skewed study data. For example, patients with lower income may be less likely to seek healthcare, and patients with higher income may be more likely to have mental health conditions screened for.

Methods & Study Design

Last, we will discuss limitations of the methods we chose for this study. First, a 30-day prediction window is chosen for this study based on clinical input; however, we do not test alternate time windows for their potential impacts on performance. Similarly, we do not test input time windows other than 90 days. These time window setpoints could be negatively affecting model performance; however, we found only minor differences in performance between high and low healthcare-utilization cohorts. Second, we use synthetic oversampling[98] to generate additional suicide attempt records during model training. Synthetic oversampling helps to combat class imablance[96]; however, it may also amplify biases present in the underlying data. Finally, we rely on the Vanderbilt WordcloudIndexer[52], a regular expressions[50] algorithm, to extract medical concepts from clinical notes. The indexer was designed to recognize and ignore negations[52]; however, the program may still fail to detect unexpected text patterns, like typos.

Future work

Natural Language Processing

In the future, we plan to apply more advanced NLP methods to further improve our risk prediction model. For example, we believe that a vector embedding step, such as word2vec[53] or cui2vec[54], may prove beneficial. Word2Vec uses skip-gram negative sampling to learn vector-space embeddings for words in the corpus. Cui2Vec uses a combination of embedding techniques and transfer learning to create vector embeddings for UMLS CUIs. These vector embedding methods use cooccurrences to transform terms into vectors which carry added information, such as similarity to other terms in the corpus. Vector embedding has proven beneficial in the past, especially for deep-learning models.[53,56] These studies usually perform transfer learning of pretrained embedding models on ordered-texts representative of their corpus before transforming their data. One challenge to this future aim is that we do not have easy access to ordered texts to perform transfer learning on.

We have generated a few ideas to overcome this limitation. Prior studies on suicidality and NLP have used texts mined from social media platforms to perform their transfer learning on.[114,115] We may be able to apply a similar approach; however, there could be limited overlap between our medical vocabulary and social media vocabulary. Other word embedding techniques have used non-ordered 'global' contexts (i.e., document-level cooccurrences) for model training.[116] Though workable in theory, global embeddings are highly computationally demanding and may prove infeasible with our resources. Our final idea is to obtain access to a smaller number of ordered clinical note texts and perform transfer learning on a limited subset of CUIs which we have already identified in feature importance studies. These restrictions on dataset size and corpus length should significantly reduce the resource demands of transfer learning a vector embedding model. Of course, it is likely that our future work will include a hybrid approach based on all three of these ideas.

Validation

In addition, we would like to further validate our model performance. Retrospective models can be prone to overfitting and model drift; thus, prospective validation is an important step to show that retrospective models can achieve clinically meaningful performance going forward. VSAIL was initially trained and validated on a retrospective dataset.[23] Later, Walsh et. al. performed prospective validation to show VSAIL's clinical relevance as a predictive screening tool.[41] Thus, we aim to follow a similar prospective validation as was done by Walsh et. al. with VSAIL. Another important future step is to show external validity with our model. External validation helps show that a model is well-fit to universal underlying patterns and is not overly impaired by selection or information biases arising from the institution where it was trained. Together, these added validation studies would strengthen the case for using our model as a clinical suicide risk prediction tool.

Clinical Deployment

Finally, we would like to take steps toward clinical deployment of our model. Successful clinical deployment of a machine learning model is a highly challenging endeavor, which requires stakeholder collaboration, application development, validation, and maintenance. We would like to follow the clinical deployment steps taken by Walsh et. al. with VSAIL.[41] However, in this study there are additional complexities to consider with backend application development, such as how best to aggregate and access CUI counts. Fortunately, the present deployment efforts with VSAIL will help to blaze the trail for our fusion models, which are still a while out in development and validation. Streamliner, another VUMC project, is a clinical application which displays structured suicide risk factors present in a patient's chart. One major benefit of Streamliner is that it can be used to shed light onto the causative factors influencing VSAIL's predictions. It may also be prudent to integrate clinical note features from our model into Streamliner to provide deeper insights into a patient's risk for suicide. We are hopeful that we will be able to validate our clinical note / structured data hybrid suicide risk prediction model and deploy it in the clinical setting for maximal impact.

Appendix 1: Role of the student in the manuscripts

Chapter 1: Introduction

Research and literature review by KK. Authoring by KK with revision assistance from committee (CW, SD, ZY).

Chapter 2: Network Analysis

Study design developed by KK with assistance from PI CW. KK performed literature review with assistance from JS. KK performed all computational & quantitative analyses – including data preprocessing, network inference, and plotting. KK and JS analyzed results to formulate discussion and conclusion. KK authored chapter with first-draft assistance from AB and JS. KK performed heavy individual edits to prepare manuscript for submission to AMIA, with review by CW, SD, and ZY.

Chapter 3: Supervised Learning

Study design developed by KK with assistance from PI CW and committee (SD, ZY). Analyses performed by KK. Chapter written by KK. Review and revision by CW, SD, and ZY.

Chapter 4: Summary & Conclusion

Research and literature review by KK. Authoring by KK with revision assistance from committee (CW, SD, ZY).

Appendix 2: Bibliography

1. Stone, D. M., Jones, C. M. & Mack, K. A. Changes in Suicide Rates — United States, 2018–2019. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 261–268 (2021).
2. Ma, J., Ward, E. M., Siegel, R. L. & Jemal, A. Temporal Trends in Mortality in the United States, 1969-2013. *JAMA* **314**, 1731 (2015).
3. *RePORTER: 'Suicide' Project Funding through Time*. https://reporter.nih.gov/search/kQIaZYSuOUmrqL83zDa7nA/projects/charts?shared=true.
4. Sym, J. *Lifes preservative against self-killing*. (Routledge, 2015).
5. Goldney, R. D., Schioldann, J. A. & Dunn, K. I. Suicide Research before Durkheim. *Health Hist.* **10**, 73 (2008).
6. Moore, C., Moore, C. & Moore, C. *A full inquiry into the subject of suicide: to which are added two treatises on duelling and gaming*. (Thoemmes Press, 1998).
7. Mueller, A. S., Abrutyn, S., Pescosolido, B. & Diefendorf, S. The Social Roots of Suicide: Theorizing How the External Social World Matters to Suicide and Suicide Prevention. *Front. Psychol.* **12**, 621569 (2021).
8. Durkheim, E. *Suicide*. (Routledge, 2005). doi:10.4324/9780203994320.
9. Joiner, T. E. *Why people die by suicide*. (Harvard University Press, 2005).
10. Van Orden, K. A. *et al.* The interpersonal theory of suicide. *Psychol. Rev.* **117**, 575–600 (2010).
11. Klonsky, E. D. & May, A. M. The Three-Step Theory (3ST): A New Theory of Suicide Rooted in the "Ideation-to-Action" Framework. *Int. J. Cogn. Ther.* **8**, 114–129 (2015).
12. Klonsky, E. D., Saffer, B. Y. & Bryan, C. J. Ideation-to-action theories of suicide: a conceptual and empirical update. *Curr. Opin. Psychol.* **22**, 38–43 (2018).
13. Klonsky, E. D., May, A. M. & Saffer, B. Y. Suicide, Suicide Attempts, and Suicidal Ideation. *Annu. Rev. Clin. Psychol.* **12**, 307–330 (2016).
14. O'Connor, R. C. & Kirtley, O. J. The integrated motivational–volitional model of suicidal behaviour. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170268 (2018).
15. Deisenhammer, E. A. *et al.* The duration of the suicidal process: how much time is left for intervention between consideration and accomplishment of a suicide attempt? *J. Clin. Psychiatry* **70**, 19–24 (2009).
16. Weissman, M. M. *et al.* Prevalence of suicide ideation and suicide attempts in nine countries. *Psychol. Med.* **29**, 9–17 (1999).
17. Zalsman, G. *et al.* Suicide prevention strategies revisited: 10-year systematic review. *Lancet Psychiatry* **3**, 646–659 (2016).
18. Mann, J. J. *et al.* Suicide prevention strategies: a systematic review. *JAMA* **294**, 2064–2074 (2005).
19. Brown, G. K. *et al.* Cognitive Therapy for the Prevention of Suicide Attempts: A Randomized Controlled Trial. *JAMA* **294**, 563 (2005).
20. McCauley, E. *et al.* Efficacy of Dialectical Behavior Therapy for Adolescents at High Risk for Suicide: A Randomized Clinical Trial. *JAMA Psychiatry* **75**, 777 (2018).

21. Linehan, M. M. *et al.* Two-Year Randomized Controlled Trial and Follow-up of Dialectical Behavior Therapy vs Therapy by Experts for Suicidal Behaviors and Borderline Personality Disorder. *Arch. Gen. Psychiatry* **63**, 757 (2006).

22. Posner, K. *et al.* The Columbia–Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *Am. J. Psychiatry* **168**, 1266–1277 (2011).

23. Walsh, C. G., Ribeiro, J. D. & Franklin, J. C. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clin. Psychol. Sci.* **5**, 457–469 (2017).

24. Bejan, C. & Walsh, C. G. Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. (in revision).

25. Zhang, D., Yin, C., Zeng, J., Yuan, X. & Zhang, P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med. Inform. Decis. Mak.* **20**, 280 (2020).

26. Spokas, M., Wenzel, A., Brown, G. K. & Beck, A. T. Characteristics of individuals who make impulsive suicide attempts. *J. Affect. Disord.* **136**, 1121–1125 (2012).

27. Meyer, I. H., Blosnich, J. R., Choi, S. K., Harper, G. W. & Russell, S. T. Suicidal Behavior and Coming Out Milestones in Three Cohorts of Sexual Minority Adults. *LGBT Health* **8**, 340–348 (2021).

28. Miranda-Mendizabal, A. *et al.* Gender differences in suicidal behavior in adolescents and young adults: systematic review and meta-analysis of longitudinal studies. *Int. J. Public Health* **64**, 265–283 (2019).

29. Parascandola, M. Causation in epidemiology. *J. Epidemiol. Community Health* **55**, 905–912 (2001).

30. Dickerman, B. A. & Hernán, M. A. Counterfactual prediction is not only for causal inference. *Eur. J. Epidemiol.* (2020) doi:10.1007/s10654-020-00659-8.

31. Franklin, J. C. *et al.* Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol. Bull.* **143**, 187–232 (2017).

32. Pelton, M. *et al.* Rates and risk factors for suicidal ideation, suicide attempts and suicide deaths in persons with HIV: a systematic review and meta-analysis. *Gen. Psychiatry* **34**, e100247 (2021).

33. Boggs, J. M. *et al.* General Medical, Mental Health, and Demographic Risk Factors Associated With Suicide by Firearm Compared With Other Means. *Psychiatr. Serv. Wash. DC* **69**, 677–684 (2018).

34. Beghi, M., Rosenbaum, J. F., Cerri, C. & Cornaggia, C. M. Risk factors for fatal and nonfatal repetition of suicide attempts: a literature review. *Neuropsychiatr. Dis. Treat.* **9**, 1725–1736 (2013).

35. Levis, M., Leonard Westgate, C., Gui, J., Watts, B. V. & Shiner, B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol. Med.* **51**, 1382–1391 (2021).

36. de Beurs, D. Network Analysis: A Novel Approach to Understand Suicidal Behaviour. *Int. J. Environ. Res. Public. Health* **14**, 219 (2017).

37. Simons, J. S. *et al.* Nexus of despair: A network analysis of suicidal ideation among veterans. *Arch. Suicide Res.* **24**, 314–336 (2020).

38. Graziano, R. C. *et al.* A network analysis of risk factors for suicide in Iraq/Afghanistan-era veterans. *J. Psychiatr. Res.* **138**, 264–271 (2021).

39. Gijzen, M. W. M. *et al.* Suicide ideation as a symptom of adolescent depression. a network analysis. *J. Affect. Disord.* **278**, 68–77 (2021).

40. Bloch-Elkouby, S. *et al.* The suicide crisis syndrome: A network analysis. *J. Couns. Psychol.* **67**, 595–607 (2020).

41. Walsh, C. G. *et al.* Prospective Validation of an Electronic Health Record–Based, Real-Time Suicide Risk Model. *JAMA Netw. Open* **4**, e211428 (2021).

42. Hong, N. *et al.* Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2017**, 74–83 (2018).

43. Dolin, R. H. *et al.* The HL7 Clinical Document Architecture. *J. Am. Med. Inform. Assoc.* **8**, 552–569 (2001).

44. OMOP Common Data Model.

45. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).

46. Gandomi, A. & Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **35**, 137–144 (2015).

47. Boggs, J. M., Quintana, L. M., Powers, J. D., Hochberg, S. & Beck, A. Frequency of Clinicians' Assessments for Access to Lethal Means in Persons at Risk for Suicide. *Arch. Suicide Res.* **26**, 127–136 (2022).

48. Zhong, Q.-Y. *et al.* Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. *Eur. J. Epidemiol.* **34**, 153–162 (2019).

49. McCoy, T. H., Castro, V. M., Roberson, A. M., Snapper, L. A. & Perlis, R. H. Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. *JAMA Psychiatry* **73**, 1064 (2016).

50. Thompson, K. Programming Techniques: Regular expression search algorithm. *Commun. ACM* **11**, 419–422 (1968).

51. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).

52. Mandani, S., Giuse, D., McLemore, M. & Weitkamp, A. Augmenting NLP Results byLeveraging SNOMED CT Relationships forIdentification of Implantable CardiacDevices from Patient Notes. (2019).

53. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs* (2013).

54. Beam, A. L. *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *ArXiv180401486 Cs Stat* (2019).

55. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

56. Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**, i37–i48 (2017).

57. Ng, A. Y. Preventing 'Overfitting' of Cross-Validation Data. in *Proceedings of the Fourteenth International Conference on Machine Learning* 245–253 (Morgan Kaufmann Publishers Inc., 1997).

58. Fu, S. *et al.* Clinical Concept Extraction: a Methodology Review. *J. Biomed. Inform.* **109**, 103526 (2020).

59. Turney, P. D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *arXiv:cs/0212032* (2002).

60. Dey, L. & Haque, S. K. M. Opinion mining from noisy text data. in *Proceedings of the second workshop on Analytics for noisy unstructured text data - AND '08* 83–90 (ACM Press, 2008). doi:10.1145/1390749.1390763.

61. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

62. Euler, L. Solutio problematis ad geometriam situs pertinentis. *Comment. Acad. Sci. Imp. Petropolitanae* **8**, 128–140 (1736).

63. Invariants. in *Graph Theory* 1–20 (John Wiley & Sons, Ltd, 2000). doi:https://doi.org/10.1002/9781118033043.ch1.

64. Moreno, J. L. *Who shall survive?: A new approach to the problem of human interrelations.* (Nervous and Mental Disease Publishing Co, 1934). doi:10.1037/10648-000.

65. Brandes, U. & Erlebach, T. Introduction. in *Network Analysis: Methodological Foundations* (eds. Brandes, U. & Erlebach, T.) 1–6 (Springer Berlin Heidelberg, 2005). doi:10.1007/978-3-540-31955-9_1.

66. Ebadi, A., Tighe, P. J., Zheng, L. & Rashidi, P. A Quest for the Structure of Intra- and Postoperative Surgical Team Networks: Does the Small World Property Evolve over Time? *ArXiv180303359 Cs* (2018).

67. McCurdie, T., Sanderson, P. & Aitken, L. M. Applying social network analysis to the examination of interruptions in healthcare. *Appl. Ergon.* **67**, 50–60 (2018).

68. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

69. Oates, Chris. J. & Mukherjee, S. Network inference and biological dynamics. *Ann. Appl. Stat.* **6**, (2012).

70. Hevey, D. Network analysis: a brief overview and tutorial. *Health Psychol. Behav. Med.* **6**, 301–328 (2018).

71. McCallum, A. K., Nigam, K., Rennie, J. & Seymore, K. Automating the Construction of Internet Portals with Machine Learning. *Inf. Retr.* **3**, 127–163 (2000).

72. Brinkmeier, M. & Schank, T. Network Statistics. in *Network Analysis: Methodological Foundations* (eds. Brandes, U. & Erlebach, T.) 293–317 (Springer Berlin Heidelberg, 2005). doi:10.1007/978-3-540-31955-9_11.

73. Armour, C., Fried, E. I., Deserno, M. K., Tsai, J. & Pietrzak, R. H. A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *J. Anxiety Disord.* **45**, 49–59 (2017).

74. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. in *Proceedings of the 23rd international conference on Machine learning - ICML '06* 161–168 (ACM Press, 2006). doi:10.1145/1143844.1143865.

75. Altman, N. & Krzywinski, M. Simple linear regression. *Nat. Methods* **12**, 999–1000 (2015).

76. Slinker, B. K. & Glantz, S. A. Multiple Linear Regression: Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable. *Circulation* **117**, 1732–1737 (2008).

77. Nick, T. G. & Campbell, K. M. Logistic regression. *Methods Mol. Biol. Clifton NJ* **404**, 273–301 (2007).

78. Ng, A. Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. in *Twenty-first international conference on Machine learning - ICML '04* 78 (ACM Press, 2004). doi:10.1145/1015330.1015435.

79. World Health Organization. *Preventing suicide: a global imperative*. (World Health Organization, 2014).

80. Hedegaard, H. *et al.* Issues in Developing a Surveillance Case Definition for Nonfatal Suicide Attempt and Intentional Self-harm Using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) Coded Data. *Natl. Health Stat. Rep.* 1–19 (2018).

81. Bloch, F., Jackson, M. O. & Tebaldi, P. Centrality Measures in Networks. *ArXiv160805845 Phys.* (2021).

82. Taylor, D., Myers, S. A., Clauset, A., Porter, M. A. & Mucha, P. J. Eigenvector-Based Centrality Measures for Temporal Networks. *ArXiv150701266 Nlin Physicsphysics* (2016).

83. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proceedings of the 7th Python in Science Conference* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (2008).

84. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).

85. Oji, C., Moore, T. A. & Gutierrez, C. A. A review of electroconvulsive therapy in suicidality. *Ment. Health Clin.* **5**, 212–215 (2015).

86. Liu, R. T., Kleiman, E. M., Nestor, B. A. & Cheek, S. M. The Hopelessness Theory of Depression: A Quarter Century in Review. *Clin. Psychol. Publ. Div. Clin. Psychol. Am. Psychol. Assoc.* **22**, 345–365 (2015).

87. Bernert, R. A., Kim, J. S., Iwata, N. G. & Perlis, M. L. Sleep disturbances as an evidence-based suicide risk factor. *Curr. Psychiatry Rep.* **17**, 554 (2015).

88. Pigeon, W. R., Pinquart, M. & Conner, K. Meta-analysis of sleep disturbance and suicidal thoughts and behaviors. *J. Clin. Psychiatry* **73**, e1160-1167 (2012).

89. Magrabi, F. *et al.* Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb. Med. Inform.* **28**, 128–134 (2019).

90. Garcia-Vidal, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E. & Soriano, A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine* **46**, 27–29 (2019).

91. Noorbakhsh-Sabet, N., Zand, R., Zhang, Y. & Abedi, V. Artificial Intelligence Transforms the Future of Health Care. *Am. J. Med.* **132**, 795–801 (2019).

92. Ancker, J. S. *et al.* Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med. Inform. Decis. Mak.* **17**, 36 (2017).

93. Sparck Jones, K. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *J. Doc.* **28**, 11–21 (1972).

94. Guyon, I. & Elisseeff, A. An Introduction to Variable and Feature Selection. *J Mach Learn Res* **3**, 1157–1182 (2003).

95. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv201016061 Cs Stat* (2020).

96. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259 (2018).

97. Zhang, J. & Mani, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. in *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets* (2003).

98. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

99. Berkson, J. Application to the Logistic Function to Bio-Assay. *J. Am. Stat. Assoc.* **39**, 357 (1944).

100. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 785–794 (2016) doi:10.1145/2939672.2939785.

101. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

102. Demir-Kavuk, O., Kamada, M., Akutsu, T. & Knapp, E.-W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinformatics* **12**, 412 (2011).

103. Ghosh, S. & Rudy, Y. Application of L1-norm regularization to epicardial potential solution of the inverse electrocardiography problem. *Ann. Biomed. Eng.* **37**, 902–912 (2009).

104. Ogunleye, A. & Wang, Q.-G. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 2131–2140 (2020).

105. Platt, J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv Large Margin Classif* **10**, (2000).

106. Spiegelhalter, D. J. Probabilistic prediction in patient management and clinical trials. *Stat. Med.* **5**, 421–433 (1986).

107. Ross, E. L. *et al.* Accuracy Requirements for Cost-effective Suicide Risk Prediction Among Primary Care Patients in the US. *JAMA Psychiatry* **78**, 642 (2021).

108. Scornet, E. Trees, forests, and impurity-based variable importance. Preprint at http://arxiv.org/abs/2001.04295 (2021).

109. Nimon, K. F. Statistical Assumptions of Substantive Analyses Across the General Linear Model: A Mini-Review. *Front. Psychol.* **3**, (2012).

110. St»hle, L. & Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **6**, 259–272 (1989).

111. Tripepi, G., Jager, K. J., Dekker, F. W. & Zoccali, C. Selection Bias and Information Bias in Clinical Research. *Nephron Clin. Pract.* **115**, c94–c99 (2010).

112. Gabella, B. A., Hume, B., Li, L., Mabida, M. & Costich, J. Multi-site medical record review for validation of intentional self-harm coding in emergency departments. *Inj. Epidemiol.* **9**, 16 (2022).

113. Jonsson Funk, M. & Landi, S. N. Misclassification in Administrative Claims Data: Quantifying the Impact on Treatment Effect Estimates. *Curr. Epidemiol. Rep.* **1**, 175–185 (2014).

114. Thiruvalluru, R. K., Gaur, M., Thirunarayan, K., Sheth, A. & Pathak, J. Comparing Suicide Risk Insights derived from Clinical and Social Media data. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2021**, 364–373 (2021).

115. Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I. & Reichart, R. Deep neural networks detect suicide risk from textual facebook posts. *Sci. Rep.* **10**, 16685 (2020).

116. Pennington, J., Socher, R. & Manning, C. Glove: Global Vectors for Word Representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014). doi:10.3115/v1/D14-1162.