Spring 2022

Deep Risk Prediction and Embedding of Patient Data: Application to Acute Gastrointestinal Bleeding

Dennis Shung
Yale University Graduate School of Arts and Sciences, dennis.shung@gmail.com

Abstract

Deep Risk Prediction and Embedding of Patient Data:

Application to Acute Gastrointestinal Bleeding

Dennis Legen Shung

2022

Acute gastrointestinal bleeding is a common and costly condition, accounting for over 2.2 million hospital days and 19.2 billion dollars of medical charges annually. Risk stratification is a critical part of initial assessment of patients with acute gastrointestinal bleeding. Although all national and international guidelines recommend the use of risk-assessment scoring systems, they are not commonly used in practice, have sub-optimal performance, may be applied incorrectly, and are not easily updated.

With the advent of widespread electronic health record adoption, longitudinal clinical data captured during the clinical encounter is now available. However, this data is often noisy, sparse, and heterogeneous. Unsupervised machine learning algorithms may be able to identify structure within electronic health record data while accounting for key issues with the data generation process: measurements missing-not-at-random and information captured in unstructured clinical note text. Deep learning tools can create electronic health record-based models that perform better than clinical risk scores for gastrointestinal bleeding and are well-suited for learning from new data. Furthermore, these models can be used to predict risk trajectories over time, leveraging the longitudinal nature of the electronic health record.

The foundation of creating relevant tools is the definition of a relevant outcome measure; in acute gastrointestinal bleeding, a composite outcome of red blood cell transfusion, hemostatic intervention, and all-cause 30-day mortality is a relevant, actionable outcome that reflects the need for hospital-based intervention. However, epidemiological trends may affect the relevance and effectiveness of the outcome measure when applied across multiple settings and patient populations.

Understanding the trends in practice, potential areas of disparities, and value proposition for using risk stratification in patients presenting to the Emergency Department with acute gastrointestinal bleeding is important in understanding how to best implement a robust, generalizable risk stratification tool. Key findings include a decrease in the rate of red blood cell transfusion since 2014 and disparities in access to upper endoscopy for patients with upper gastrointestinal bleeding by race/ethnicity across urban and rural hospitals. Projected accumulated savings of consistent implementation of risk stratification tools for upper gastrointestinal bleeding total approximately $1 billion 5 years after implementation.

Most current risk scores were designed for use based on the location of the bleeding source: upper or lower gastrointestinal tract. However, the location of the bleeding source is not always clear at presentation. I develop and validate electronic health record based deep learning and machine learning tools for patients presenting with symptoms of acute gastrointestinal bleeding (e.g., hematemesis, melena, hematochezia), which is more relevant and useful in clinical practice. I show that they outperform leading clinical risk scores for upper and lower gastrointestinal bleeding, the Glasgow Blatchford Score and the Oakland score. While the best performing gradient boosted decision tree model has equivalent overall performance to the fully connected feedforward neural network model, at the very low risk threshold of 99% sensitivity the deep learning model identifies more very low risk patients. Using another deep learning model that can model longitudinal risk, the long-short-term memory recurrent neural network, need for transfusion of red blood cells can be predicted at every 4-hour interval in the first 24 hours of intensive care unit stay for high-risk patients with acute gastrointestinal bleeding.

Finally, for implementation it is important to find patients with symptoms of acute gastrointestinal bleeding in real time and characterize patients by risk using available data in the electronic health record. A decision rule-based electronic health record phenotype has equivalent performance as measured by positive predictive value compared to deep learning and natural language processing-based models, and after live implementation appears to have increased the use of the Acute Gastrointestinal Bleeding Clinical Care pathway. Patients with acute gastrointestinal bleeding but with other groups of disease concepts can be differentiated by directly mapping unstructured clinical text to a common ontology and treating the vector of concepts as signals on a knowledge graph; these patients can be differentiated using unbalanced diffusion earth mover’s distances on the graph. For electronic health record data with data missing not at random, MURAL, an unsupervised random forest-based method, handles data with missing values and generates visualizations that characterize patients with gastrointestinal bleeding.

This thesis forms a basis for understanding the potential for machine learning and deep learning tools to characterize risk for patients with acute gastrointestinal bleeding. In the future, these tools may be critical in implementing integrated risk assessment to keep low risk patients out of the hospital and guide resuscitation and timely endoscopic procedures for patients at higher risk for clinical decompensation.

Deep Risk Prediction and Embedding of Patient Data:

Application to Acute Gastrointestinal Bleeding

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Dennis Legen Shung

Dissertation Director: Smita Krishnaswamy

May 2022

Table of Contents

Dedication

Wir sollen beten, vertrauen, und von ihm alles annehmen.

Martin Luther, Jakobus 4,7 Lutherbibel 1912

I dedicate the work, which has taken me beyond what I have ever asked for or imagined, to the infinite grace of God, the source of Lux et Veritas.

Acknowledgements

I am deeply, profoundly grateful to my life partner and wife Stephanie, whose sacrificial love and struggle made my work possible. To Aidan and Julia, my light and joy. To my parents, David and Doris, who encouraged me to persevere despite having no experience with academia or medicine. To Loren, whose incisive mind and exacting standards have sharpened my mind and taught me how to do things excellently. To Smita, who has pushed me to achieve things that I did not think were possible, and who has brought me through the enlightening process of algorithmic development. To Perry and Allen, who have helped hone my theoretical dreams with the reality of informatics in practice. To Gene, whose tough love and grants workshop have resulted in funding to pursue and deepen my research career. To Fred, who encouraged me to consider the Ph.D. and has been an incredible supporter throughout the process. To Kenneth, whose patient and excellent instruction helped nurture a fledgling interest in machine learning and who has been a true friend. To Michael, whose enthusiasm and parsimonious coding have powered me through the initial vagaries of SQL data loading, and whose friendship I have drawn upon in times of difficulty. To Joseph Sung, whose conversations have inspired me to persevere in academia. To Alex, who has been so generous with his time in helping me understand fundamental concepts and get excited about future work together. Finally, to all of my collaborators, colleagues, mentors, and friends who have made this journey worthwhile: thank you.

Introduction

Gastrointestinal bleeding (GIB) is the most common cause of GI-related hospitalization in the United States (U.S) and accounts for $19.2 billion in hospital-related charges with over 500,000 hospitalizations annually.[1] Risk stratification has the potential to decrease unnecessary hospitalizations. With the advent of care delivery through the electronic health records, there is an exciting new opportunity to access and analyze data generated through clinical care. However, electronic health record data is messy, sparse, and varied, primarily generated for billing purposes and therefore not optimized for computation with traditional statistical tools. New tools should be considered that adapt to specific aspects of the electronic health record data generation process. As a physician-scientist, I have sought to tackle the challenge of using electronic health record data using novel machine learning and deep learning tools while also being firmly grounded in the clinical value proposition that leads to adoption of such tools on a provider and health system level. This thesis seeks to 1) define the trends, possible sources of bias, and value proposition of machine learning approaches for risk stratification in gastrointestinal bleeding; 2) develop and validate EHR-based tools for identifying patients with EHR phenotyping and providing risk stratification to identify very low risk patients who could be discharged directly from the ED with machine learning and deep learning tools; 3) explore dynamic risk prediction for patients who have changes in risk or who are at high risk; 4) create new tools tailored for the challenges of using EHR data to characterize patient cohorts and provide continuously updated risk prediction.

Artificial intelligence (AI) is a broad term that encompasses a diverse array of subfields. Machine learning, a subset of artificial intelligence, refers to a group of computational tools, or algorithms, that can be trained to learn specific patterns within data and optimize prediction.

Machine learning has the potential to enhance the practice of medicine.[2] However, an 'AI chasm' has been described that limit the clinical application of machine learning models.[3] Clinicians are domain experts that can help bridge the gap by becoming active partners in developing and implementing machine learning models for clinical use. The paradigm of collaboration between domain experts and machine learning engineers has been successful in developing expert-augmented machine learning.[4] Gastrointestinal bleeding (GIB) is the most common cause of GI-related hospitalization in the United States (U.S), accounting for 19.2 billion USD of estimated charges.[1]

The future of gastrointestinal bleeding will include the integration of machine learning algorithms to enhance clinician risk assessment and decision-making. For risk assessment, the goals are twofold: first, triage to the appropriate level of care, and second, to inform decisions for testing and treatment.

Triage of very low risk patients for outpatient management rather than hospitalization is recommended by national and international guidelines for both upper and lower gastrointestinal bleeding.[5-9] The benefit for identifying high risk patients is uncertain, but has been explored recently by predicting in-hospital mortality in patients presenting with acute gastrointestinal bleeding admitted to the intensive care unit.[10]

Machine learning tools have specific advantages over clinical risk scores, including the ability to improve over time and to be retrained with data reflecting local epidemiology and patterns of disease.

Upper Gastrointestinal Bleeding:

Existing clinical risk scores use statistical models applied to a mixture of clinical, laboratory, and demographic data taken during initial assessment and have been shown to 1) accurately identify very low risk patients and 2) decrease unnecessary healthcare utilization when applied consistently. More importantly, two prospective trials have shown that consistent application of the Glasgow Blatchford Score at very low risk thresholds (score 0 or less than 1) significantly decreases hospital admissions (96% without using the GBS to 71% with GBS;P<0.001, 94% without using GBS versus 82% with using GBS;P<0.001).[11,12] Machine learning scores appear to perform better than clinical risk scores, and in particular outperforms the GBS at the very low risk threshold.[13] A gradient-boosted machine learning model outperformed the most common and validated clinical risk scores for acute upper GIB, the admission Rockall, AIMS65, and GBS in predicting a composite outcome reflecting need for hospitalization (packed red blood cell transfusion, hemostatic intervention, or 30-day mortality) with significantly increased overall are under the receiver operating curve and at the very low risk cutoff.[14]

Lower Gastrointestinal Bleeding:

The Oakland score has been shown to perform well in identifying very low risk patients with lower gastrointestinal bleeding, and also has been externally validated in an electronic health record database.[15,16] However, there are no prospective trials applying the score to test its performance in affecting differences in healthcare utilization. Machine learning scores, particularly neural network models, appear to perform better than clinical risk scores for patients with lower gastrointestinal bleeding.[17-19]

A key challenge to applying these scores for risk stratification has been provider uptake and application in clinical practice. Only about 30% of all physician providers in the United States have ever used a score in caring for patients with acute upper gastrointestinal bleeding.[20] The reasons for this may include ignorance, cumbersomeness of calculating the score, and unclearly defined responsibility in the use of these risk scores.[21] For lower gastrointestinal bleeding, none of the scores developed have been adopted consistently in clinical practice, and existing scores suffer from limitations of being developed in large administrative data registries and in small samples of patients.[22]

Electronic Health Records: A New Era of Learning Health Systems

With the advent of electronic health records deployed across the world, there is a clear opportunity to address the challenge of deployment of risk stratification tools, particularly machine learning tools. Electronic health records provide infrastructure that can be leveraged to deploy these tools in real time. The potential of data streams generated through the electronic health record can be harnessed to identify patients with GIB, provide timely risk stratification based on presenting vital signs, laboratory results, and co-morbidities, and then monitor ongoing risk to guide decisions regarding clinical care.[23] Machine learning tools are well-suited for identifying patients using EHR-based phenotypes and developing prognostic risk models with electronic health record data.[24,25]

Phenotyping is the first step to leveraging the EHR for risk stratification in acute GIB

Identifying patients with a specific condition or outcome is a central challenge to deploy risk stratification tools in the EHR. For acute gastrointestinal bleeding, an automated mechanism that would identify patients at the time of presentation would be the first step in consistently deploying risk stratification tools.

For EHR phenotyping, and two primary approaches have been deployed: expert-driven decision rules and machine learning approaches. The challenge comes from the mixture of datatypes in EHR data and the dynamic nature of data collection. These datatypes are broadly defined as structured datafields, or discrete values that are stored individually in specified categories including lab values and diagnosis codes, and unstructured datafields, user generated blocks of data that do not fall under specific categories such as clinician notes. For expert-driven decision rules, experts define the condition through the presence of specific conditions to include or exclude patients (e.g., laboratory exam fecal calprotectin for inflammatory bowel disease, ICD code for reflux disease) and then use these conditions to create decision rules. Machine learning models include both supervised and unsupervised learning models, which can be coupled with data preprocessing to extract information from unstructured text through natural language processing. In particular, unsupervised machine learning approaches have been used to identify variables without the need for expert-defined conditions, and have demonstrated promise in providing portability across different medical centers.[26-28]

For gastrointestinal bleeding, the relevant population of patients are patients who present with either reported symptoms of overt gastrointestinal bleeding or provider observed signs of active gastrointestinal bleeding. Inherently this can only be approximated by diagnosis codes entered during a hospital encounter, since not all patients receive therapy (inpatient endoscopic procedures, blood product transfusions) and the lab tests used to assess risk are nonspecific (e.g., hemoglobin, INR). In a large cohort of patients presenting with possible symptoms of gastrointestinal bleeding, we have conducted the first study to test and compare the performance of expert decision rules and natural language processing approaches to identify patients presenting with symptoms of overt acute gastrointestinal bleeding.[29]

Limitations include interpretability, ability to incorporate unstructured data reliably, and generalizability. While decision rule models are simple and interpretable, they usually cannot use unstructured data, have poor performance and do not generalize easily across practice settings.[30-32] Machine learning models are promising but have not undergone extensive validation studies. For all phenotyping efforts, the major limitation is the generation of expert-labeled datasets, which are time intensive and difficult to replicate.

EHR-based Risk Prediction Models: Static and Dynamic

Clinical risk scores can and should be mapped to EHR datafields to enable automated calculation. The key challenge is understanding how to translate the specific components of the risk score to structured datafields captured in the EHR. For example, the Oakland score for acute lower GIB has been translated to an “EHR-friendly” format, which necessitated the removal of the digital rectal exam and definition of “previous lower GIB” through a combination of ICD-10 diagnosis codes.[16]

Machine learning risk prediction tools developed using EHR data have shown promise in when deployed to predict sepsis, acute kidney injury, and in-hospital delirium.[33-36] However, there has been limited work in acute GIB.

One study using EHR data to develop machine learning models in acute gastrointestinal bleeding uses a gradient boosted tree machine learning that is superior to the APACHE score in predicting inpatient mortality risk for high risk patients admitted to the intensive care unit for 24 hours.[10]

While initial risk assessment is helpful for identifying low risk patients who could undergo outpatient management, dynamic continuous risk assessment can leverage the longitudinal information in the EHR to guide triage and may prove useful for anticipating clinical deterioration or timing therapeutic interventions such as packed red blood cell transfusion. The cornerstone of management of patients with acute GIB is adequate resuscitation using intravenous fluids and packed red blood cell transfusion.

Overview of Material Covered in This Work:

I start with a detailed epidemiological study of trends in acute gastrointestinal bleeding, including components of the composite outcome, to evaluate for changes in practice patterns and types of GIB. Furthermore, I evaluate for disparities in access to endoscopic evaluation for patients presenting with UGIB by race/ethnicity and rural versus urban, teaching versus non-teaching hospitals. I define the value of consistent implementation of risk stratification systems for UGIB using a cost-minimization analysis.

I then present findings using structured EHR data from the first 4 hours to develop and validate machine learning and deep learning models. I present work to develop EHR phenotypes that can be used in real time to identify patients with acute GIB using structured and unstructured data with preliminary data from real-time implementation of the decision rule phenotype to suggest the use of a Clinical Care pathway for Acute Gastrointestinal Bleeding.

I present a dynamic risk modeling approach using Long-Short Term Memory Recurrent Neural Networks to predict need for red blood cell transfusion in high-risk patients with GIB admitted to the ICU.

Finally, I created a new algorithmic approach to visualize electronic health record data with data missing-not-at-random and present a proof of concept of using standard clinical ontologies to map unstructured data into a standardized concept dictionary, and then applying the concepts as signals on a knowledge graph with an algorithmic innovation to separate different diagnostic phenotypes. Future work will include thinking through issues of data access, bias, equity, and challenges to implementation in healthcare systems.

Key terms:

Artificial intelligence (AI): Generally, the ability for a computer to accomplish tasks typically associated with human intelligence.

Machine learning (ML): a subfield of artificial intelligence, broadly refers to the ability of a computational platform to learn from data and make predictions or recommendations based on this data without being explicitly programmed

Supervised learning is conducted with the concept of “truth” where the model tries to approximate the relationship between inputs and labeled outputs. For example, given images of cats and dogs, where each image has a correct answer, can you train a model that accurately identifies of cats versus dogs?

Unsupervised learning is performed without data labels and the goal is for the computer to infer inherent structure or patterns in the data. For example, given a set of heart rate, accelerometer, and location data from a wearable fitness monitor, can the computer identify periods of rest versus exercise based on differences in the raw data?

Neural networks (NN): a form of machine learning with a basic architecture consisting of nodes and connections existing in multiple layers, loosely analogous to neurons and synapses in the biological brain. This broad category is inclusive of many kinds of modern machine learning models which are used in tasks such as computer vision, voice recognition, bioinformatics, among others.

Deep learning: A broad family of neural network architectures that have multiple layers (aka deep).

Epidemiology of Acute Gastrointestinal Bleeding in the Emergency Department

Introduction

Gastrointestinal bleeding (GIB) is the most common cause of GI-related hospitalization in the United States (U.S).[1] Traditionally, epidemiologic studies have classified GIB into upper GIB (UGIB) and lower GIB (LGIB) based on the location of the bleeding source. UGIB refers to bleeding from the esophagus, stomach, or duodenum, and LGIB arises from the small and large bowel distal to the Ligament of Treitz.

Previous epidemiological studies have reported a low case fatality rate for GIB in the U.S. (<5%) and suggested decreasing incidence, but these studies only evaluated hospitalized patients and did not include patients who presented and were discharged from emergency departments (ED).[37-41] Moreover, the most recent studies that reported on GIB trends over time evaluated patients no later than 2012 for UGIB and 2014 for LGIB.[37-41] Possible factors contributing to the downward trend in UGIB incidence include Helicobacter pylori eradication and the increased use of proton pump inhibitors.[42] The most recent epidemiological study of LGIB in the U.S. suggested an increased hospitalization rate for LGIB since 2010,[40] thought to be related to an aging population and increasing frequency of antithrombotic therapy.[40,42] Furthermore, new evidence regarding management of GIB and resultant changes in guideline recommendations (e.g., restrictive red blood cell [RBC] transfusion strategies, discharge of very-low-risk patients from the ED for outpatient management) may have affected clinical practice patterns.

We believe that an updated and comprehensive epidemiological evaluation of the incidence and secular trends, as well as hospital-based management and outcomes, for all patients with GIB presenting to EDs in the U.S. is needed to better understand the characteristics and current state of care of patients with GIB. This study uses a large, national emergency department database to characterize epidemiological trends for GIB incidence, management, and outcomes – and assesses whether real-world management has changed with the advent of new guideline recommendations.

Materials and Methods

Data source:

This study includes data from 2006 to 2019 collected in the Nationwide Emergency Department Sample (NEDS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality.[43] NEDS is the largest all-payer ED database in the U.S and is a composite of the HCUP State Inpatient Database and the State Emergency Department Database, which includes information on ED visits that result in admission at that ED's hospital and those that do not, respectively. NEDS includes information on patients treated and discharged from the ED, those seen in the ED and admitted to the same hospital, and those who were transferred from the ED to another hospital. For patients who were transferred, only information from their ED visit was available. The composite database from 2019 describes an unweighted sample of 33.1 million discrete ED visits across 989 hospitals in 41 states, including the District of Columbia, and a weighted national estimate of 143.4 million ED visits. Weighted samples were used for the purposes of this study. We performed final checks to comply with HCUP privacy protections policy and confirm that no individual persons have been identified either directly or indirectly, hospitals have not been identified, with all aggregate statistical reporting containing at least two hospitals in any individual cell, and no cell sizes are included that are less than or equal to 10.

Variables:

We identified ED visits with a primary diagnosis of GIB using the International Classification of Disease (ICD) codes (**Table 1**).

Table 1: ICD-9-CM and ICD-10-CM codes for gastrointestinal bleeding

	ICD-9-CM	**ICD-10-CM ***
Upper GIB		

Bleeding Ulcer	530.21	K22.11,
	531.00, 531.01, 531.20,	K25.0, K25.2, K25.4, K25.6,
	531.21, 531.40, 531.41,	K26.0, K26.2, K26.4, K26.6,
	531.60, 531.61,	K27.0, K27.2, K27.4, K27.6,
	532.00, 532.01, 532.20,	K28.0, K28.2, K28.4, K28.6
	532.21, 532.40, 532.41,	
	532.60, 532.61,	
	533.00, 533.01, 533.20,	
	533.21, 533.40, 533.41,	
	533.60, 533.61,	
	534.00, 534.01, 534.20,	
	534.21, 534.40, 534.41,	
	534.60, 534.61	
Variceal Bleed	456.0, 456.20	I85.01, I85.11
Mallory-Weiss Tears	530.7	K22.6
Angiodysplasia	537.83	K31.811
Dieulafoy’s Lesion	537.84	K31.82
Gastritis/Duodenitis	535.01, 535.11, 535.21,	K29.01, K29.21, K29.31,
	535.31, 535.41, 535.51,	K29.41, K29.51, K29.61,
	535.61, 535.71	K29.71, K29.81, K29.91
Other	530.82, 578.0	K22.8, K92.0

Lower GIB		
Diverticular Bleed	562.02, 562.03, 562.12, 562.13	K57.01, K57.11, K57.13, K57.21, K57.31, K57.33, K57.41, K57.51, K57.53, K57.81, K57.91, K57.93
Anorectal Hemorrhage	569.3	K62.5
Hemorrhoid	455.0, 455.1, 455.2, 455.5, 455.6, 455.8	K64.0, K64.1, K64.2, K64.3, K64.4, K64.8
Angiodysplasia	569.85	K55.21
Dieulafoy’s Lesion	569.86	K63.81
Other	569.41, 565.0	K62.6, K60.2, K62.7, K51.411
Unspecified GIB	578.1, 578.9	K92.1, K92.2, K91.840

* Other ICD-10-CM codes for GIB have been added after 2019 (e.g., K21.01, K20.81, K20.91) and therefore were not present in our dataset.

Due to the shift from the ninth revision (ICD-9) to tenth revision (ICD-10) after the third quarter of 2015, we used ICD-9 from 2006 to 2015 3rd quarter and ICD-10 from 2015 4th quarter to 2019. Diagnosis of GIB were categorized into upper, lower, and unspecified GIB with further specification of the types of UGIB and LGIB. Types of UGIB included bleeding from ulcers, varices, Mallory-Weiss tears, angiodysplasia, Dieulafoy’s lesions, gastritis or duodenitis, and other non-specific UGIB diagnoses (e.g., unspecified esophageal hemorrhage, hematemesis). Types of LGIB included diverticular bleed, anorectal hemorrhage, angiodysplasia, Dieulafoy's lesions, hemorrhoids, and other intestinal and anorectal causes (e.g., bleeding from polyp, anorectal ulcer). Unspecified GIB included blood in stool defined as melena or hematochezia, unspecified hemorrhage of the gastrointestinal tract, and unspecified postprocedural gastrointestinal hemorrhage.

Demographic data included sex, age, geography based on zip code (Northeast, Midwest, South, West), income quartile based on estimated median household income of residents in the patient's zip code, and primary insurance payer (Medicare, Medicaid, private insurance, self-pay, no charge, other). Charlson Comorbidity Index (CCI) scores, which predict 10-year survival in patients with multiple comorbidities, were calculated based on ICD codes as described by Glasheen et. al with minor revisions.[44]

Study Outcomes:

The outcomes of interest were incidence of GIB, rates of RBC transfusions, case fatality, ED discharge, presence of upper or lower endoscopy, average inpatient length of stay, and inpatient healthcare costs. Age- and sex-adjusted incidence and rates were calculated to remove the confounding effects of changes in age and sex composition across time. We included patients aged ≥20 years in our analyses as population structure estimates for age and sex were only available in 5-year intervals from the U.S. Census Bureau (e.g., age=15-19, 20-25) and GIB is rare in individuals aged <20 years.[45]

For incidence, we calculated crude incidence rate for each calendar year by dividing the total number of ED visits with a primary diagnosis of GIB by the U.S. population estimate for the respective year. We used the 2010 U.S. population structure as the standard to calculate age- and sex-adjusted incidence rate by summing the multiplicative product of each age- and sex-specific incidence by the age- and sex-specific standard population divided by the total standard population. We similarly calculated age-adjusted incidence rates for GIB stratified by sex and stratified by 10-

year age group categories. 10-year age group stratified data from 2006 was not available since the U.S. Census Bureau did not report uniform age categories in 2006 for age population structure.

Rates of RBC transfusion, case fatality, ED discharge, and upper or lower endoscopy were standardized to the respective GIB population structure in 2010 following the method described above. RBC transfusions and endoscopic procedures were identified using codes from both Current Procedure Terminology (CPT), a set of medical codes used to report medical, surgical, and diagnostic procedures and services, and ICD Procedure Coding System (ICD-PCS) (**Table 2**). Case fatality rate was defined as the proportion of all-cause, in-facility deaths among the total ED visits with a primary diagnosis of GIB. The specific cause of death was not available in this database.

Table 2: CPT and ICD-PCS codes for transfusions and endoscopies.

	CPT	**ICD-9-PCS**	**ICD-10-PCS**
RBC Transfusion	36430, 36440, 36450, 36444, 36455, 36456, 36460	99.00, 99.02, 99.03, 99.04	30233N0, 30233N1, 30233H0, 30233H1, 30233P0, 30233P1, 30243N0, 30243N1, 30243H0, 30243H1, 30243P0, 30243P1, 30230N0, 30230N1, 30230H0, 30230H1, 30230P0, 30230P1, 30240N0, 30240N1, 30240H0, 30240H1, 30240P0, 30240P1

Upper endoscopy	*Diagnostic* 43235, 43239 *Any hemostasis* 43255 *Submucosal injection* 43236, 43243 *Band ligation* 43244	*Diagnostic* 42.21, 42.22, 42.23, 42.24, 44.11, 44.12, 44.13, 44.14, 45.11, 45.11, 45.13, 45.14, 45.16, *Esophagus hemostasis* 42.33, 42.82 *Stomach hemostasis* 44.43 *Duodenum hemostasis* 45.30 *Submucosal injection* 39.92 *Band ligation* 42.91	*Diagnostic* 0DJ04ZZ, 0DJ08ZZ, 0DJ64ZZ, 0DJ68ZZ, 0DB14ZX, 0DB24ZX, 0DB34ZX, 0DB44ZX, 0DB54ZX, 0DB64ZX, 0DB74ZX, 0DB84ZX, 0DB94ZX, 0DB18ZX, 0DB28ZX, 0DB38ZX, 0DB48ZX, 0DB58ZX, 0DB68ZX, 0DB78ZX, 0DB88ZX, 0DB98ZX *Esophagus hemostasis* 0D514ZZ, 0D524ZZ, 0D534ZZ, 0D544ZZ, 0D554ZZ, 0D518ZZ, 0D528ZZ, 0D538ZZ, 0D548ZZ, 0D558ZZ, 0DQ14ZZ, 0DQ24ZZ, 0DQ34ZZ, 0DQ44ZZ, 0DQ54ZZ, 0DQ18ZZ, 0DQ28ZZ, 0DQ38ZZ, 0DQ48ZZ, 0DQ58ZZ

			Stomach hemostasis 0D564ZZ, 0D574ZZ, 0D568ZZ, 0D578ZZ, 0DQ64ZZ, 0DQ74ZZ, 0DQ68ZZ, 0DQ78ZZ *Duodenum hemostasis* 0D594ZZ, 0D598ZZ, 0DQ84ZZ, 0DQ94ZZ, 0DQ88ZZ, 0DQ98ZZ *Unspecified upper GI hemostasis* 3E0G8TZ *Band ligation* 06L34CZ, 06L38CZ
Lower endoscopy	*Diagnostic* 45378, 45300, 45330, *Any hemostasis* 45382, 45317, 45334,	*Diagnostic* 45.21, 45.22, 45.23, 45.24, 45.25 48.21, 48.22, 48.23, 48.24 *Any hemostasis*	*Diagnostic* 0DJD4ZZ, 0DJD8ZZ, 0DBA4ZX, 0DBB4ZX, 0DBC4ZX, 0DBE4ZX, 0DBF4ZX, 0DBG4ZX, 0DBH4ZX, 0DBK4ZX,

	45.43, 48.32	0DBL4ZX, 0DBM4ZX,
Submucosal		0DBN4ZX, 0DBP4ZX,
injection	*Submucosal injection*	0DBA8ZX, 0DBB8ZX,
45381, 45335,	39.92	0DBC8ZX, 0DBE8ZX,
		0DBF8ZX, 0DBG8ZX,
Band ligation	*Hemorrhoid ligation*	0DBH8ZX, 0DBK8ZX,
45398, 45350	49.45	0DBL8ZX, 0DBM8ZX,
		0DBN8ZX, 0DBP8ZX
		Any hemostasis
		0D5A4ZZ, 0D5B4ZZ,
		0D5C4ZZ, 0D5E4ZZ,
		0D5F4ZZ, 0D5G4ZZ,
		0D5H4ZZ, 0D5K4ZZ,
		0D5L4ZZ, 0D5M4ZZ,
		0D5N4ZZ, 0D5P4ZZ,
		0D5A8ZZ, 0D5B8ZZ,
		0D5C8ZZ, 0D5E8ZZ,
		0D5F8ZZ, 0D5G8ZZ,
		0D5H8ZZ, 0D5K8ZZ,
		0D5L8ZZ, 0D5M8ZZ,
		0D5N8ZZ, 0D5P8ZZ,
		0DQA4ZZ, 0DQB4ZZ,
		0DQC4ZZ, 0DQE4ZZ,
		0DQF4ZZ, 0DQG4ZZ,
		0DQH4ZZ, 0DQK4ZZ,
		0DQL4ZZ, 0DQM4ZZ,
		0DQN4ZZ, 0DQP4ZZ,

	0DQA8ZZ, 0DQB8ZZ, 0DQC8ZZ, 0DQE8ZZ, 0DQF8ZZ, 0DQG8ZZ, 0DQH8ZZ, 0DQK8ZZ, 0DQL8ZZ, 0DQM8ZZ, 0DQN8ZZ, 0DQP8ZZ *Hemorrhoid ligation* 06LY4CC, 06LY8CC
Unspecified location of endoscopy	*Unspecified hemostasis* 0W3P4ZZ, 0W3P8ZZ

Inpatient length of stay was a provided variable in NEDS. We used two methods to evaluate inpatient healthcare costs (reported in US dollars). First, we converted the NEDS inpatient charge variable to inpatient cost using the HCUP hospital-level charge-to-cost ratios dataset, which was available for years 2012 to 2019.[46] Second, we converted NEDS provided Medicare Severity Diagnosis Related Group (DRG) to average national Medicare payment using the Medicare Inpatient Hospitals dataset, which was available for years 2014 to 2019.[47] Both NEDS-derived inpatient cost and Medicare DRG-derived cost exclude provider fees. The NEDS-derived cost will be referred to as "inpatient cost" and the Medicare DRG-derived cost will be referred to as "Medicare payment" hereafter. Following guidelines for inflation adjustments,[48] we adjusted both calculations of inpatient healthcare costs to 2019 using the Gross Domestic Product index.[49]

Statistical analysis:

Data extraction and preprocessing was performed using Python 3.8.5 (Python Software Foundation, Wilmington, DE, USA). Statistical analyses were performed using R v4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) and STATA v17.0 (StataCorp LLC., College Station, TX, USA). Simple linear regressions were employed to analyze linear trends over 2006 to 2019 for age- and sex-adjusted GIB incidence and rates of RBC transfusions, case fatality, ED discharge, and upper or lower endoscopy. Univariate and multivariate weighted logistic regressions were performed to determine if there was a relationship between the year of GIB ED visits and outcomes (RBC transfusion, case fatality, discharge from ED, and endoscopic evaluation). Trends in inpatient length of stay and inpatient healthcare costs were evaluated using a weighted negative binomial regression and log-linked gamma generalized linear model, respectively. Multivariate models were adjusted for age (per 5-year increase), sex, hospital region, income quartile, and CCI score. Patients older than 95 years were excluded from the model due to a small sample size. All statistical tests were based on two-tailed probability.

Results

The total ED visits with GIB as the primary diagnosis from the NEDS database was 172,358 in 2006 and 241,077 in 2019, with the national weighted projected cases of 804,604 in 2006 and 1,043,604 in 2019 (**Table 1**). The distributions of age, sex, hospital region, and income quartile for GIB cases were similar across the 14-year study period. For primary payor, more patients were on Medicaid in 2019 (15.0%) than 2006 (9.0%). GIB cases with greater comorbidities (CCI ≥4) increased from 8.9% in 2006 to 15.6% in 2019 with corresponding decreases in cases with minimal or no comorbidities (CCI ≤1: 2006=73.1%, 2019=64.4%).

Table 1. Selected characteristics for gastrointestinal bleeding patients from the National Emergency Department Sample, 2006 – 2019.

Characteristics	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Total (n)	804604	812410	837528	850863	870232	894664	892278	914763	945515	997371	1006088	1038566	1018495	1043604
Location of Bleed (%)														
Upper	28.8	27.7	26.3	25.8	25.6	25.1	24.6	24.7	23.7	24.0	23.9	25.2	27.2	28.7
Lower	40.4	40.6	40.4	41.0	41.4	42.5	42.4	42.0	42.1	41.0	39.6	39.3	38.6	38.5
Unspecified	30.8	31.7	33.3	33.2	33.0	32.4	33.0	33.3	34.2	35.0	36.5	35.5	34.2	32.8
Female (%)	48.8	49.3	49.1	49.0	48.7	48.9	48.5	48.3	48.2	48.3	48.0	47.8	47.6	47.5
Age, year (%)														
< 25	5.0	5.1	5.2	5.5	5.6	5.8	5.8	5.4	5.5	5.3	5.3	4.7	4.5	4.6
25-44	19.9	20.1	20.3	20.0	20.6	20.0	20.2	19.5	19.8	20.0	20.5	19.8	19.5	19.7
45-64	26.2	26.4	26.8	27.3	27.6	27.3	27.8	27.3	27.9	27.7	28.0	27.9	27.4	27.4
≥ 65	49.0	48.5	47.6	47.3	46.2	46.9	46.2	47.8	46.7	47.0	46.3	47.6	48.6	48.3
Region (%)														
Northeast	19.3	19.4	19.0	18.4	18.5	19.0	18.6	18.7	17.9	17.0	17.7	18.3	18.4	18.4
Midwest	23.1	23.1	22.3	22.9	23.6	22.8	22.5	22.3	23.2	23.2	23.2	22.8	22.7	23.5
South	38.5	39.1	39.4	39.4	38.7	38.9	39.5	39.8	39.7	39.6	39.2	39.6	39.0	38.3
West	19.0	18.4	19.2	19.3	19.2	19.2	19.2	19.2	19.2	20.0	19.8	19.3	19.8	19.8
Income Quartile (%)														
$1 - 43,999	29.3	30.0	28.7	29.1	29.7	28.7	30.4	30.9	31.6	32.3	31.5	30.8	32.4	31.0
$44,000 - 55,999	25.0	25.9	28.3	27.7	26.6	25.3	25.3	26.6	28.4	24.1	26.7	26.3	27.0	25.2
$56,000 - 73,999	23.5	23.2	21.2	22.6	22.5	24.2	23.1	22.7	21.0	23.4	21.7	22.7	21.0	23.0
$74,000+	20.1	18.6	19.1	18.1	19.0	19.6	19.2	17.5	16.8	18.4	18.3	18.6	18.0	19.1
Primary Payor (%)														
Medicare	51.4	50.2	49.6	49.3	48.9	50.1	49.7	51.1	50.0	50.2	49.4	50.3	51.1	50.4
Medicaid	9.0	9.3	10.0	10.7	11.6	11.9	12.3	12.1	15.8	15.7	15.8	15.4	15.7	15.0
Private Insurance	24.8	25.5	25.7	24.6	23.2	22.5	21.5	21.5	21.8	23.0	23.8	23.0	22.4	23.3
Self-Pay	10.6	11.4	11.1	11.3	12.2	11.5	12.3	11.2	9.0	7.7	7.8	8.2	8.0	8.3
No Charge	0.9	0.6	0.7	0.9	0.7	0.7	0.7	1.0	0.6	0.5	0.4	0.5	0.3	0.4
Other	3.0	2.7	2.8	2.9	3.1	3.1	3.4	3.1	2.8	2.9	2.8	2.4	2.4	2.6
Charlson Co-morbidity Index (%)														
0	51.5	50.7	50.4	49.8	49.8	49.0	49.5	47.9	47.9	47.5	48.2	46.8	45.5	45.0
1	21.6	21.9	21.6	21.3	21.4	21.4	21.1	21.3	20.6	20.5	19.9	19.9	19.8	19.4
2	11.2	12.3	12.1	12.3	12.2	12.4	12.1	12.6	12.1	11.9	11.2	11.2	11.2	11.1
3	6.8	7.5	7.7	7.8	7.9	8.1	8.1	8.4	8.4	8.4	8.3	8.4	8.7	8.8
4+	8.9	7.7	8.1	8.7	8.8	9.1	9.2	9.8	11.0	11.7	12.3	13.8	14.8	15.6

The age- and sex-adjusted incidence for ED visits for GIB increased from 387.9/100,000 population in 2006 to 407.1/100,000 in 2019 (P-value <0.001; **Figure 1A**). The increase in incidence was larger in male patients (411.7/100,000 in 2006, 438.5/100,000 in 2019; P <0.001) than in female patients (2006 = 365.5/100,000, 2019 = 377.6/100,000; P=0.021). UGIB incidence decreased from 112.3/100,000 in 2006 to 94.4/100,000 in 2014 but then increased to 116.2/100,000 by 2019 (**Figure 1B**). In contrast, incidence for LGIB increased from 155.5/100,000 in 2006 to a peak of 171.9/100,000 in 2015 but then declined to 159.8/100,000 by 2019. Incidence for unspecified GIB increased from 120.1/100,000 in 2006 to a peak of 148.1/100,000 in 2016 before then declining to 131.2/100,000 in 2019. We observed similar trends for UGIB, LGIB, and unspecified when stratifying by male and female patients.

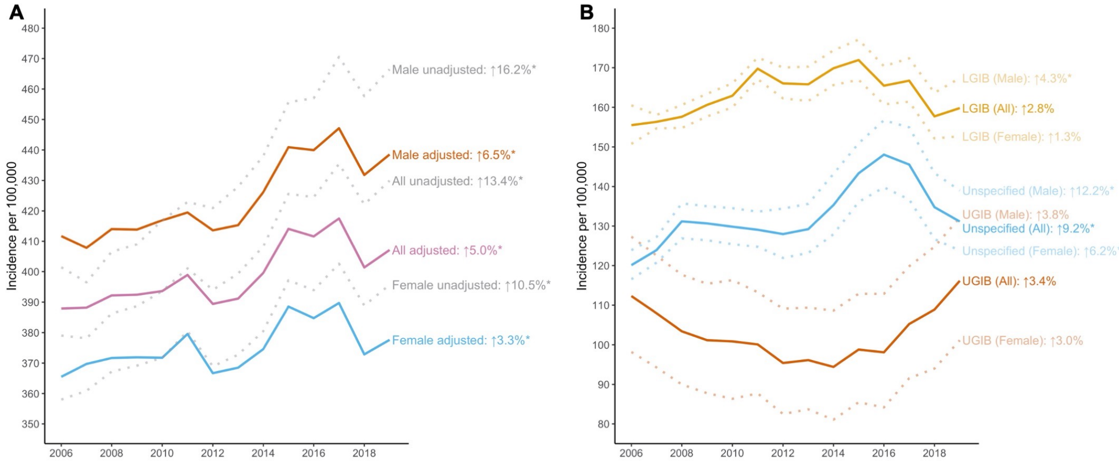


Figure 1. Incidences (per 100,000 population) and percent change from 2006 to 2019 for gastrointestinal bleeding (GIB) as the primary diagnosis for emergency department visit stratified by male, female, and all (A) and stratified by upper, lower, and unspecified GIB (B). Percentages indicate overall percent change from 2006 to 2019. *P<0.05 for trend.

Cumulative percent changes in GIB incidence compared to 2007 and stratified by 10-year age categories are shown in **Figure 2**. Data from 2006 was not available since the U.S. Census Bureau did not report uniform categories in 2006 for age population structure. From 2007 to 2019, we observed marked increases in any GIB incidence in younger and middle age groups (**Figure 2A**). There were upward trends for age groups 20-29 years (+13.1%; P=0.033), 30-39 years (+27.5%; P <0.001), 40-49 years (+16.4%; P <0.001), and 50-59 years (+27.0%; P <0.001). However, the incidence of any GIB was relatively constant for the 60-69 years age group (+3.1%; P=0.061) and decreased for age groups of 70-79 years (-4.8%; P=0.014) and ≥80 years (-13.0%; P=0.003).

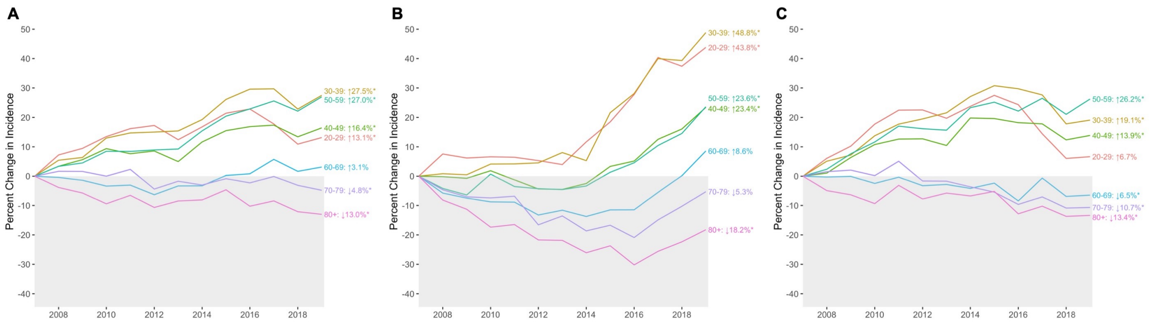


Figure 2. Cumulative percent changes in sex-adjusted incidence (per 100,000 population) compared to 2007 and stratified by categorical age for any gastrointestinal bleed (A), upper gastrointestinal bleed (B), and lower gastrointestinal bleed (C). Percentages indicate overall percent change from baseline year of 2007 to 2019. *P<0.05 for trend.

We observed similar age-stratified trends for UGIB and LGIB incidence from 2007 to 2014 with increasing incidence among younger and middle age groups but relatively stable or decreasing incidence among older patients. For UGIB (**Figure 2B**), incidence was relatively stable from 2007 to 2014 for age groups of 20-29 years, 30-39 years, 40-49 years, and 50-59 years before substantially increasing afterward, resulting in overall changes of +43.8% (P <0.001), +48.8% (P <0.001), +23.4% (P=0.004), and +23.6% (P=0.002), respectively. In contrast, UGIB incidence declined for age groups of 60-69 years and 70-79 years from 2007 to 2016 and then increased afterwards, resulting in overall changes of +8.6% (P=0.412) and -5.3% (P=0.069), respectively. UGIB incidence decreased in patients aged ≥80 years (-18.3%; P=0.002). For LGIB (**Figure 2C**), there was a consistent upward trend from 2007 to 2019 for age group 50-59 years with a percent change of +26.2% (P <0.001). For age groups 20-29 years, 30-39 years, and 40-49 years, the incidences rose from 2007 to 2015 before decreasing thereafter, resulting in overall changes of +6.7% (P=0.486), +19.1% (P=0.002), and +13.9% (P=0.002), respectively. There were decreases in patients aged 60-69 (-6.5%; P=0.004), 70-79 (-10.7%; P <0.001) and ≥80 years (-13.4%; P=0.001).

When looking at etiologies of UGIB (**Figure 3A**), incidence of bleeding from ulcers, gastritis, and angiodysplasia decreased from 2006 to 2016 and then increased afterward with overall percent changes of -8.3% (P=0.102), -2.3% (P=0.271), and +29.7% (P=0.042), respectively. Incidence of variceal bleeding was constant from 2006 to 2014 with a marked increase thereafter (overall change +73.6%; P=0.001). There was a linear decrease in Mallory-Weiss tears (-18.7%; P <0.001) and an increase in bleeding from other UGIB causes (+14.5%; P <0.001). Incidence of bleeding from Dieulafoy’s lesions was relatively constant from 2006 to 2014 (+5.4%; P=0.942).

For LGIB etiologies (**Figure 3B**), incidence of diverticular and angiodysplasia bleeding decreased from 2006 to 2016 and then increased afterwards with overall percent changes of -17.2% (P <0.001) and -5.0% (P=0.142), respectively. Incidence of bleeding from anorectal hemorrhage and hemorrhoids increased from 2006 to 2015 before decreasing thereafter with overall percent changes of +20.2% (P=0.001) and -0.7% (P=0.511), respectively. We observed an increase in LGIB from Dieulafoy's lesion (+80.8%; P <0.001). Bleeding from other LGIB causes was stable from 2006 to 2019 (+2.0%; P=0.667).

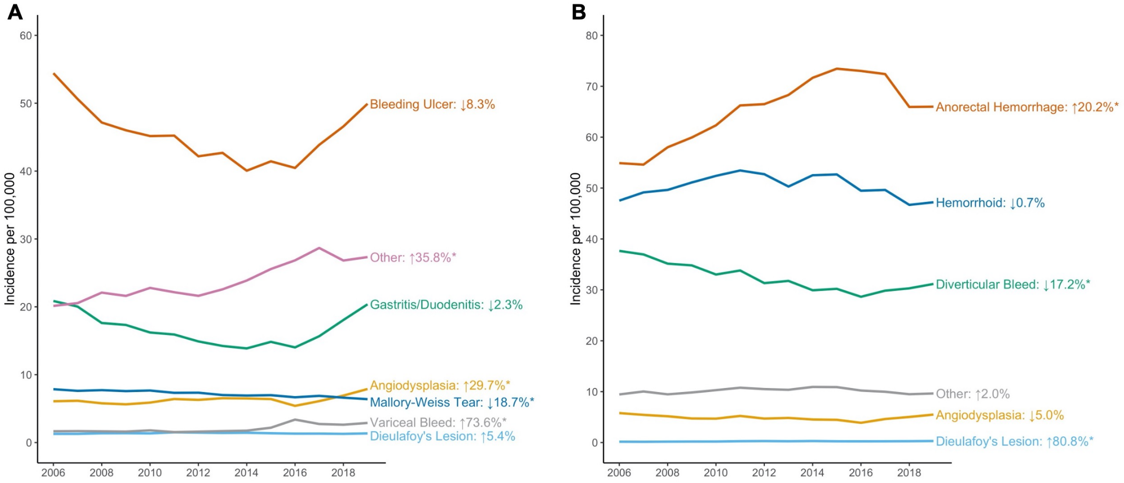
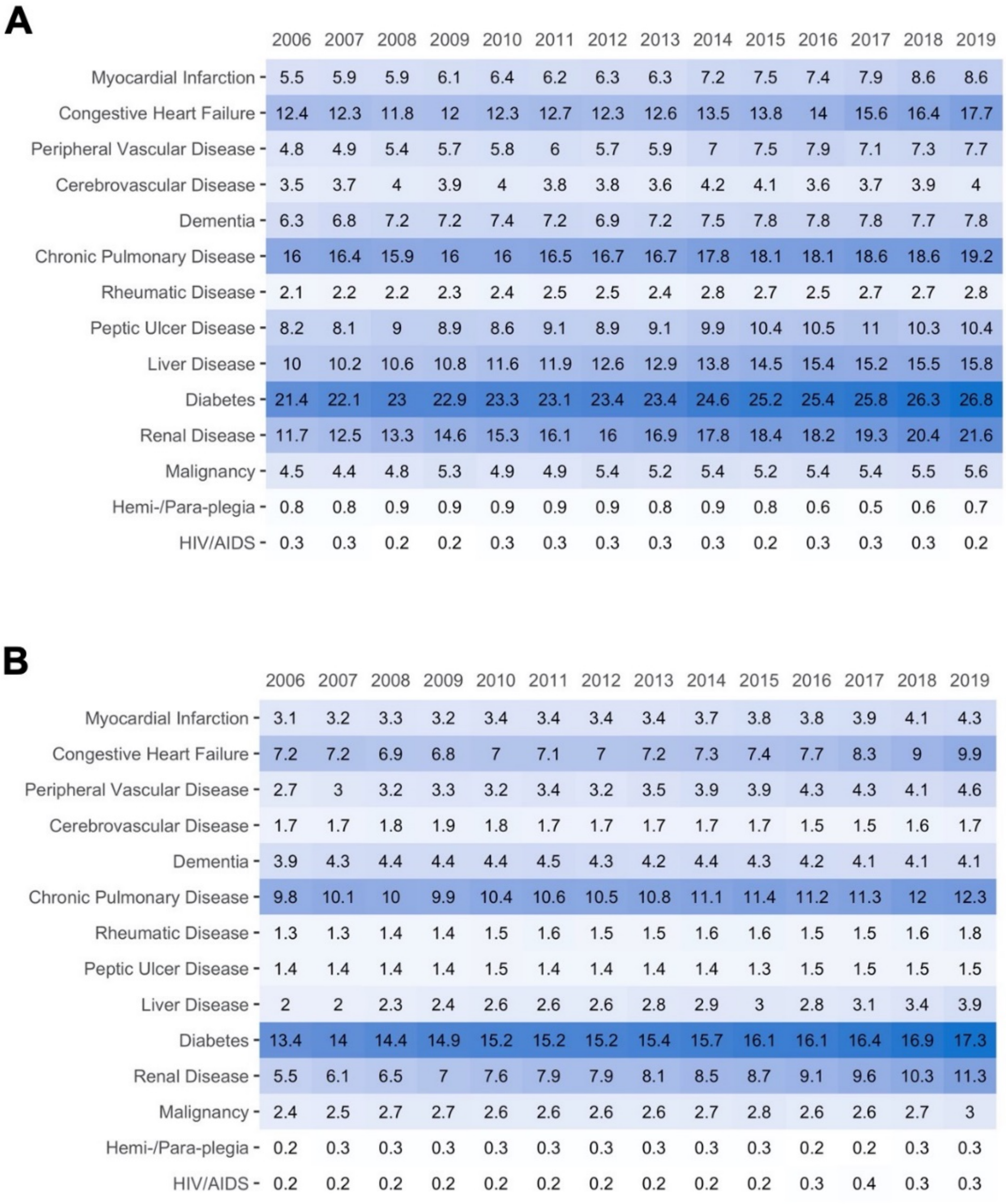


Figure 3. Age- and sex-adjusted incidences (per 100,000 population) and percent change from 2006 to 2019 for specific types of upper gastrointestinal bleed (A) and lower gastrointestinal bleed (B). Percentages indicate overall percent change from baseline year of 2006 to 2019.

With a few exceptions, proportions of UGIB patients with common comorbidities were increased from 2006 to 2019, particularly for myocardial infarction (relative change=55.2%), congestive heart failure (43.2%), peripheral vascular disease (61.3%), liver disease (58.6%), and renal disease (85.4%) (**Figure 4A**). While comorbidities for LGIB patients were fewer overall than that of UGIB patients, there were notable increases from 2006 to 2019 in myocardial infarction (42.1%), congestive heart failure (37.4%), peripheral vascular disease (70.1%), liver disease (94.8%), renal disease (103.2%), rheumatic disease (39.9%), and HIV/AIDs (55.6%) (**Figure 4B**).



A

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Myocardial Infarction	5.5	5.9	5.9	6.1	6.4	6.2	6.3	6.3	7.2	7.5	7.4	7.9	8.6	8.6
Congestive Heart Failure	12.4	12.3	11.8	12	12.3	12.7	12.3	12.6	13.5	13.8	14	15.6	16.4	17.7
Peripheral Vascular Disease	4.8	4.9	5.4	5.7	5.8	6	5.7	5.9	7	7.5	7.9	7.1	7.3	7.7
Cerebrovascular Disease	3.5	3.7	4	3.9	4	3.8	3.8	3.6	4.2	4.1	3.6	3.7	3.9	4
Dementia	6.3	6.8	7.2	7.2	7.4	7.2	6.9	7.2	7.5	7.8	7.8	7.8	7.7	7.8
Chronic Pulmonary Disease	16	16.4	15.9	16	16	16.5	16.7	16.7	17.8	18.1	18.1	18.6	18.6	19.2
Rheumatic Disease	2.1	2.2	2.2	2.3	2.4	2.5	2.5	2.4	2.8	2.7	2.5	2.7	2.7	2.8
Peptic Ulcer Disease	8.2	8.1	9	8.9	8.6	9.1	8.9	9.1	9.9	10.4	10.5	11	10.3	10.4
Liver Disease	10	10.2	10.6	10.8	11.6	11.9	12.6	12.9	13.8	14.5	15.4	15.2	15.5	15.8
Diabetes	21.4	22.1	23	22.9	23.3	23.1	23.4	23.4	24.6	25.2	25.4	25.8	26.3	26.8
Renal Disease	11.7	12.5	13.3	14.6	15.3	16.1	16	16.9	17.8	18.4	18.2	19.3	20.4	21.6
Malignancy	4.5	4.4	4.8	5.3	4.9	4.9	5.4	5.2	5.4	5.2	5.4	5.4	5.5	5.6
Hemi-/Para-plegia	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.8	0.6	0.5	0.6	0.7
HIV/AIDS	0.3	0.3	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.2

B

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Myocardial Infarction	3.1	3.2	3.3	3.2	3.4	3.4	3.4	3.4	3.7	3.8	3.8	3.9	4.1	4.3
Congestive Heart Failure	7.2	7.2	6.9	6.8	7	7.1	7	7.2	7.3	7.4	7.7	8.3	9	9.9
Peripheral Vascular Disease	2.7	3	3.2	3.3	3.2	3.4	3.2	3.5	3.9	3.9	4.3	4.3	4.1	4.6
Cerebrovascular Disease	1.7	1.7	1.8	1.9	1.8	1.7	1.7	1.7	1.7	1.7	1.5	1.5	1.6	1.7
Dementia	3.9	4.3	4.4	4.4	4.4	4.5	4.3	4.2	4.4	4.3	4.2	4.1	4.1	4.1
Chronic Pulmonary Disease	9.8	10.1	10	9.9	10.4	10.6	10.5	10.8	11.1	11.4	11.2	11.3	12	12.3
Rheumatic Disease	1.3	1.3	1.4	1.4	1.5	1.6	1.5	1.5	1.6	1.6	1.5	1.5	1.6	1.8
Peptic Ulcer Disease	1.4	1.4	1.4	1.4	1.5	1.4	1.4	1.4	1.4	1.3	1.5	1.5	1.5	1.5
Liver Disease	2	2	2.3	2.4	2.6	2.6	2.6	2.8	2.9	3	2.8	3.1	3.4	3.9
Diabetes	13.4	14	14.4	14.9	15.2	15.2	15.2	15.4	15.7	16.1	16.1	16.4	16.9	17.3
Renal Disease	5.5	6.1	6.5	7	7.6	7.9	7.9	8.1	8.5	8.7	9.1	9.6	10.3	11.3
Malignancy	2.4	2.5	2.7	2.7	2.6	2.6	2.6	2.6	2.7	2.8	2.6	2.6	2.7	3
Hemi-/Para-plegia	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.3	0.3
HIV/AIDS	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.4	0.3	0.3

Figure 4: Trends in proportion of patients with specific co-morbidity as measured by the Charlson Co-morbidity Index mapped from ICD-9 and ICD-10 CM codes.

Between 2006 and 2019, the RBC transfusion rate for any GIB decreased from 23.2% to 18.2% (relative risk reduction [RRR] -21.2%, P=0.001; **Figure 5A**). UGIB cases saw the greatest raw decrease in RBC transfusion rate from 37.4% to 28.3%. Notably, the inflection point for the decreasing trend in RBC transfusion rate was around 2014. Case fatality rate decreased from 1.44% to 0.94% (RRR -34.5%, P <0.001; **Figure 5B**), rate of discharge from ED increased from 38.7% to 41.5% (RRR +7.1%, P <0.001; **Figure 5C**), and rate of any endoscopy decreased from 43.0% to 36.6% (RRR -14.9%, P <0.001; **Figure 5D**) from 2006 to 2019.

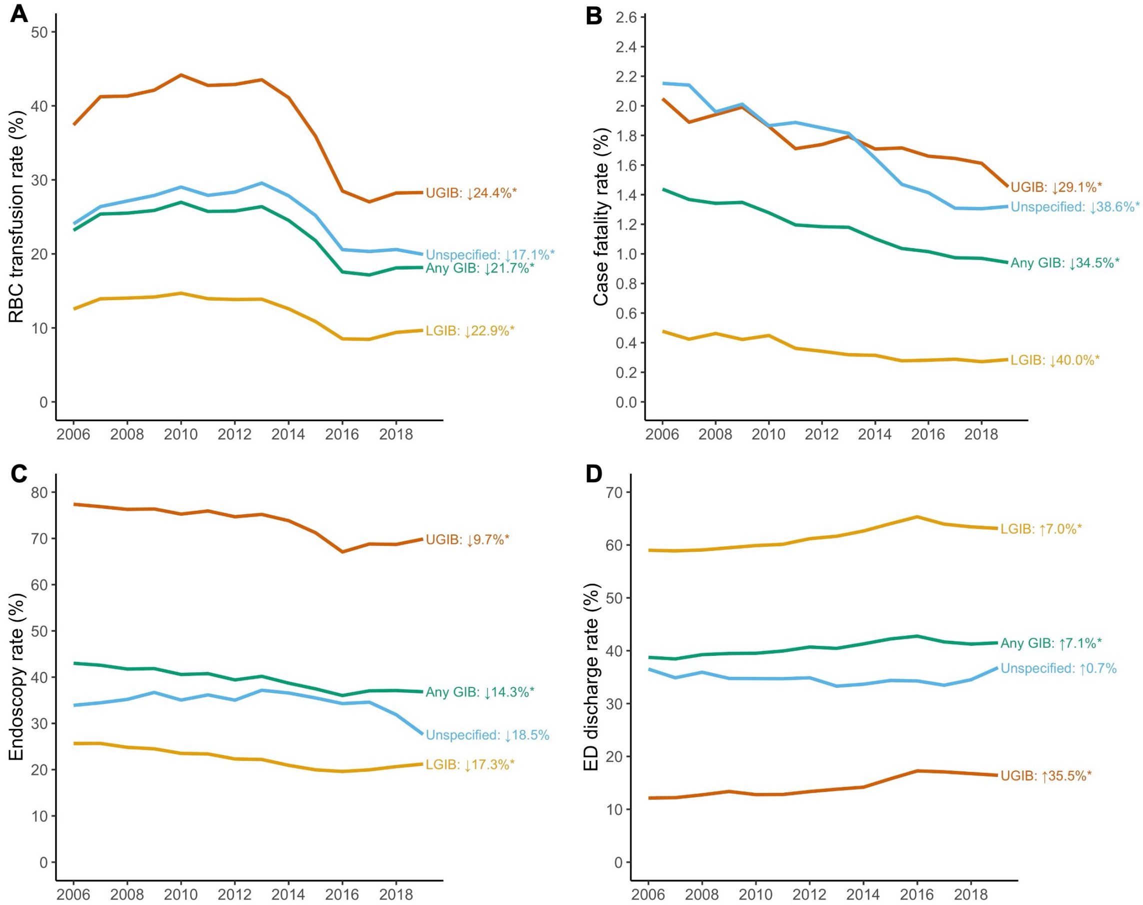


Figure 5. Age- and sex-adjusted rates and percent change from 2006 to 2019 for red blood cell transfusion (A), case fatality (B), emergency department discharge (C) and any endoscopy (D). Percentages indicate overall percent change from baseline year of 2006 to 2019. *P<0.05 for trend. RBC = red blood cell; ED = emergency department; GIB = gastrointestinal bleed; UGIB = upper GIB; LGIB = lower GIB.

Decreases in endoscopy rates for patients who were discharged from the ED (RRR -31.1%; P <0.001) were greater than for those admitted or placed under observation from the ED (RRR -7.35%; P <0.001) (**Figure 6**).

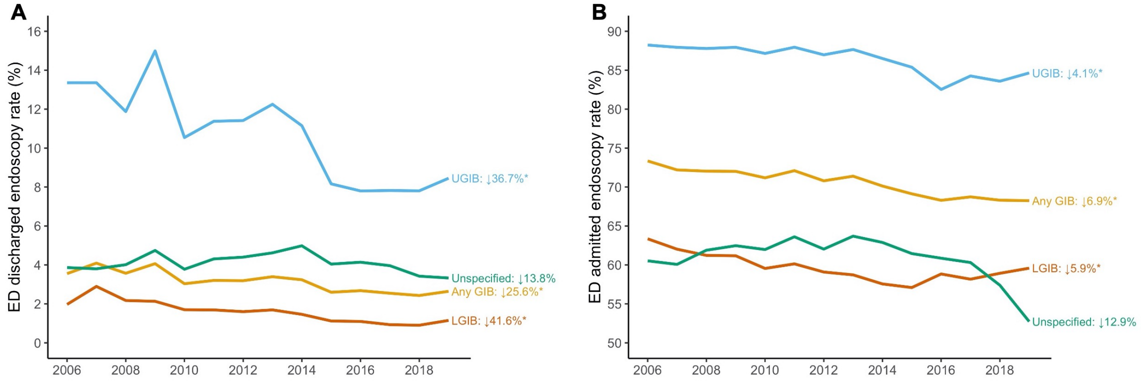


Figure 6: Trends and percentage change from 2006 to 2019 of endoscopy rate for patients discharged from the ED and those admitted for observation or inpatient stay

Results of multivariate regression analyses of outcomes by year are found in **Table 2** (univariate analyses in **Table 2A**). Compared to patients in 2006, patients in 2019 with any GIB were less likely to receive RBC transfusions (odds ratio [OR]=0.62; 95% confidence interval [CI]=0.61, 0.63) or endoscopy (OR=0.64; 95% CI=0.63, 0.65). Patients in 2019 compared to 2006 also had lower odds of death (OR=0.52; 95% CI=0.49, 0.56) and were more likely to be discharged from the ED (OR=1.44; 95% CI=1.42, 1.47). Multivariate analyses showed that patients in 2019 compared to 2006 had shorter length of stays (relative ratio [RR]=0.87; 95% CI=0.86, 0.88; **Table 2**). Inpatient healthcare costs were stable from 2012 to 2019 (RR=1.00; 95% CI=0.99, 1.01) and Medicare costs increased only slightly (RR=1.02; 95% CI=1.01, 1.02).

Table 2. Multivariate weighted regressions for gastrointestinal bleeding outcomes of interest with comparison to reference year

Year	RBC Transfusion OR (95% CI)	Case Fatality OR (95% CI)	ED Discharge OR (95% CI)	Endoscopy OR (95% CI)	Length of Stay RR (95% CI)	Inpatient Cost RR (95% CI)	Medicare Payment RR (95% C
2006	Reference	Reference	Reference	Reference	Reference		
2007	1.15 (1.13,1.17)	0.95 (0.89,1.01)	0.99 (0.98,1.01)	0.97 (0.96,0.99)	1.00 (0.99,1.01)		
2008	1.13 (1.11,1.15)	0.91 (0.86,0.97)	1.07 (1.05,1.09)	0.92 (0.90,0.93)	0.99 (0.98,1.00)		
2009	1.15 (1.13,1.17)	0.89 (0.84,0.95)	1.10 (1.08,1.12)	0.92 (0.90,0.93)	0.97 (0.96,0.98)		
2010	1.23 (1.20,1.25)	0.84 (0.79,0.90)	1.11 (1.10,1.13)	0.85 (0.84,0.87)	0.95 (0.94,0.96)		
2011	1.12 (1.10,1.14)	0.77 (0.72,0.82)	1.16 (1.14,1.18)	0.85 (0.84,0.86)	0.93 (0.92,0.94)		
2012	1.13 (1.11,1.15)	0.77 (0.73,0.82)	1.22 (1.20,1.24)	0.79 (0.78,0.80)	0.92 (0.91,0.93)	Reference	
2013	1.16 (1.14,1.18)	0.76 (0.71,0.80)	1.23 (1.20,1.25)	0.81 (0.80,0.82)	0.91 (0.90,0.92)	0.97 (0.96,0.98)	
2014	1.02 (1.00,1.04)	0.69 (0.65,0.73)	1.32 (1.29,1.34)	0.75 (0.73,0.76)	0.90 (0.89,0.91)	0.94 (0.93,0.95)	Reference
2015	0.84 (0.83,0.86)	0.65 (0.62,0.70)	1.42 (1.40,1.44)	0.69 (0.68,0.70)	0.88 (0.87,0.89)	0.93 (0.92,0.94)	1.03 (1.02,1.03)
2016	0.62 (0.60,0.63)	0.61 (0.58,0.65)	1.45 (1.43,1.48)	0.64 (0.63,0.65)	0.88 (0.87,0.89)	0.99 (0.98,1.00)	1.09 (1.08,1.09)
2017	0.59 (0.58,0.60)	0.57 (0.54,0.61)	1.40 (1.37,1.42)	0.66 (0.65,0.67)	0.87 (0.86,0.88)	0.96 (0.95,0.97)	0.98 (0.98,0.99)
2018	0.62 (0.61,0.63)	0.55 (0.52,0.58)	1.39 (1.37,1.41)	0.66 (0.65,0.67)	0.86 (0.85,0.87)	0.96 (0.95,0.98)	0.97 (0.96,0.97)
2019	0.62 (0.61,0.63)	0.52 (0.49,0.56)	1.44 (1.42,1.47)	0.65 (0.64,0.65)	0.87 (0.86,0.88)	1.00 (0.99,1.01)	1.02 (1.01,1.02)

Multivariate weighted regression adjusted for age (per 5-year increase), sex, hospital region, income quartile, and Charlson Comorbidity Index score.

RBC = red blood cell; ED = emergency department; OR = odds ratio; RR = relative ratio; CI = confidence interval.

Table 2A: Univariate weighted regressions for outcomes of interest with comparison to reference year.

Year	RBC Transfusion OR (95% CI)	Case Fatality OR (95% CI)	ED Discharge OR (95% CI)	Any Endoscopy OR (95% CI)	Length of Stay RR (95% CI)	Inpatient Cost RR (95% CI)
2006	Reference	Reference	Reference	Reference	Reference	
2007	1.12 (1.10,1.14)	0.94 (0.89,1.00)	1.00 (0.98,1.01)	0.97 (0.96,0.99)	1.01 (1.00,1.02)	
2008	1.12 (1.10,1.13)	0.91 (0.85,0.96)	1.05 (1.04,1.07)	0.93 (0.92,0.94)	0.99 (0.99,1.00)	
2009	1.13 (1.12,1.15)	0.90 (0.85,0.96)	1.07 (1.05,1.08)	0.93 (0.92,0.95)	0.98 (0.97,0.99)	
2010	1.18 (1.17,1.20)	0.85 (0.80,0.90)	1.09 (1.07,1.11)	0.87 (0.86,0.89)	0.97 (0.96,0.98)	
2011	1.12 (1.11,1.14)	0.80 (0.75,0.85)	1.09 (1.08,1.11)	0.89 (0.88,0.90)	0.95 (0.94,0.96)	
2012	1.12 (1.10,1.13)	0.78 (0.74,0.83)	1.14 (1.13,1.16)	0.83 (0.82,0.85)	0.94 (0.93,0.95)	Reference
2013	1.18 (1.16,1.20)	0.80 (0.75,0.84)	1.09 (1.08,1.11)	0.88 (0.87,0.89)	0.94 (0.93,0.95)	0.97 (0.96,0.98)
2014	1.06 (1.04,1.07)	0.74 (0.69,0.78)	1.15 (1.14,1.17)	0.82 (0.81,0.83)	0.94 (0.93,0.95)	0.96 (0.95,0.97)
2015	0.91 (0.89,0.92)	0.72 (0.68,0.77)	1.19 (1.18,1.21)	0.78 (0.77,0.79)	0.92 (0.92,0.93)	0.95 (0.94,0.97)
2016	0.69 (0.68,0.70)	0.67 (0.63,0.71)	1.23 (1.22,1.25)	0.73 (0.72,0.74)	0.93 (0.92,0.94)	1.04 (1.03,1.05)
2017	0.68 (0.67,0.69)	0.65 (0.62,0.69)	1.14 (1.13,1.16)	0.77 (0.76,0.78)	0.93 (0.92,0.94)	1.01 (1.00,1.02)
2018	0.74 (0.73,0.75)	0.65 (0.62,0.69)	1.10 (1.09,1.12)	0.78 (0.77,0.79)	0.92 (0.92,0.93)	1.03 (1.02,1.04)
2019	0.74 (0.73,0.75)	0.63 (0.60,0.67)	1.12 (1.10,1.13)	0.78 (0.77,0.79)	0.94 (0.93,0.95)	1.07 (1.06,1.09)

Multivariate weighted regression adjusted for age (per 5-year increase), sex, hospital region, inc quartile, and Charlson Comorbidity Index score.
RBC = red blood cell; ED = emergency department; OR = odds ratio; RR = relative ratio; CI = co interval.

Average inpatient length of stay for any GIB steadily decreased from 4.5 days to 4.2 days between 2006 and 2019, a relative reduction of 6.0% (P <0.001; **Figure 7A**). Patients hospitalized for UGIB had longer average inpatient stay than LGIB and unspecified GIB. Length of stay for LGIB declined from 2006 to 2015 but increased thereafter. Between 2012 and 2019, average inpatient costs increased from $6144 to $6605, a relative increase of 7.5% (P=0.043; **Figure 7B**). The average Medicare payment was $8442 in 2014 and $8849 in 2019, a relative increase of 4.8% (P=0.998; **Figure 7C**).

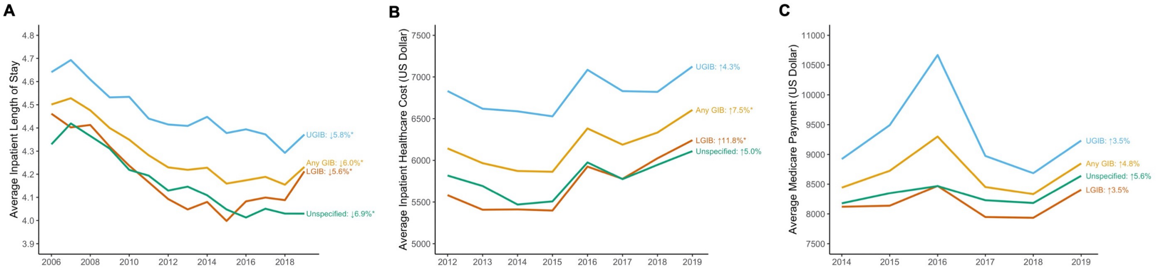


Figure 7. Trends and percent change from 2006 to 2019 for inpatient length of stay (A), inflation-adjusted average inpatient healthcare cost (B), and inflation-adjusted average Medicare payment (C). Percentages indicate overall percent change from baseline year to 2019. *P<0.05 for trend. GIB = gastrointestinal bleed; UGIB = upper GIB; LGIB = lower GIB.

Discussion

We found that the overall incidence of patients presenting to the ED with GIB in the U.S. has increased from 2006 to 2019, although changes in incidence vary based on location of GIB and age of patients. UGIB incidence has been increasing since 2014, especially in young and middle age groups (<60 years). LGIB incidence has been decreasing since 2015, but incidence in young and middle age groups (<60 years) is increasing while incidence in older age groups (≥60 years) is declining. Even though patients presenting with GIB are now sicker with more comorbidities, RBC transfusion, case fatality rate, and length of inpatient stay have decreased from 2006 to 2019, while the proportion of patients discharged from the ED has increased.

Our results regarding declining UGIB incidence from 2006 to 2014 are supported by previous studies that used national inpatient databases to examine UGIB during this timeframe.[37-39] Since the most recent trend data reported from UGIB was in 2012, the increasing incidence of UGIB after 2014 has not been shown previously. We observed increasing incidence of bleeding from ulcers, varices, gastritis/duodenitis, and angiodysplasia, suggesting that multiple causes may be driving this shift. The upward pattern after 2014 was observed as substantial increases in UGIB incidence among younger and middle-aged patients (<60 years) and smaller increases among older patients (≥60 years). Several studies showed that the expansion of Medicaid in 2014 through the Affordable

Care Act led to improved healthcare and ED utilization among younger adults.[50-53] In our study, the proportion of GIB patients with Medicaid grew by 67% between 2006 and 2019, which may have contributed to the rise in younger and middle-aged patients presenting to the ED with GIB. However, the steady increase in UGIB incidence between 2014 to 2019 across all age groups is unlikely to be explained by greater ED utilization from Medicaid expansion alone, calling for further studies on potential underlying causes for the recent rise in UGIB incidence.

The trend reversal in UGIB incidence may also be a consequence of the increasing burden of comorbidities in UGIB patients from 2006 to 2019. Increased non-gastrointestinal comorbidities comprises a strong, independent risk factor for non-variceal UGIB.[54] The growing number of patients with chronic liver disease in the U.S., especially non-alcoholic fatty liver disease, may be contributing to the increased incidence of variceal bleed.[55,56] Our study also demonstrated rising burdens of myocardial infarction, peripheral vascular disease, renal failure and congestive heart disease over time. Antithrombotic therapies, which are often prescribed for long-term prophylaxis or management of cardiovascular disease, are known risk factors for GIB.[57,58] Increasing number of patients on these antithrombotic therapies may be contributing to the more recent uptick in UGIB incidence.

Our findings of increasing LGIB incidence from 2006 to 2015 are supported by the most recent study of inpatient LGIB epidemiology from 2005 to 2014.[40] This trend in our study was greater in younger and middle-aged patients (<60 years) than in older patients (≥60 years). Decreased diverticular bleeding across that same timeframe is also consistent with other studies.[37,41] Previous studies observed that the most common inpatient diagnosis associated with LGIB was diverticular bleeding.[37,40] In contrast, we observed that anorectal hemorrhage was the most common diagnosis associated with ED visit for LGIB followed by hemorrhoidal and then diverticular bleed, likely reflecting the different patterns of LGIB etiologies in the ED versus inpatient settings. Our observation of decreasing LGIB incidence since 2015 has not been shown previously. Decreasing incidence of anorectal hemorrhage and hemorrhoids from 2015 to 2019 may be driving this trend.

Since studies of LGIB incidence have been limited, further research is needed to validate these findings.

Trends in clinical practice patterns observed in this study are consistent with updates to guideline recommendations for the management of GIB. For example, the decrease in the rate of RBC transfusion around 2014 may reflect 2010 international consensus recommendations and 2012 U.S. guideline recommendations for a restrictive transfusion strategy in patients with UGIB,[7,59] and the 2013 publication of a large randomized trial documenting improved outcomes with such a strategy.[60] Similar restrictive RBC transfusion strategies were recommended by national guidelines for LGIB in 2016.[8] The increase in ED discharge rate in our data is also consistent with updates to guideline recommendations suggesting discharge of very-low-risk patients with outpatient management.[7] We also observed declining rates of endoscopy for GIB patients. This trend is probably related to improved risk stratification and the increase in ED discharge of very-low-risk patients: patients generally do not receive endoscopy while in the ED and most of the decline in endoscopy was seen in those discharged from the ED.

Despite substantial increases in comorbidities in GIB patients presenting to the ED, case fatality continued to decrease from 2006 to 2019, which is supported by previous studies examining case fatality in both upper and lower GIB.[37-40] Similarly, average length of stay for patients hospitalized for GIB decreased from 2006 to 2019, suggesting that management and outcomes for GIB patients have steadily improved over the past two decades. Although average unadjusted NEDS-derived inpatient cost and Medicare payment showed small overall increases from 2012 to 2019 (7.5%) and 2014 to 2019 (4.8%), respectively, multivariate regression analyses showed that inpatient costs in 2019 did not change compared to 2012 and that Medicare payment in 2019 was increased only 2% compared to 2014.

Value in healthcare is defined as health outcomes achieved per dollar spent.[61] This is typically measured by evaluating improvements in quality via changes in clinical outcomes and evaluating changes in cost. A previous study using the National Inpatient Sample identified trends in in-hospital mortality and costs for bariatric surgery to suggest that increased value could be attributed to the use of new surgical techniques and technologies.[62] With GIB, we found an improvement in in-facility case fatality (either in the ED or inpatient) while maintaining roughly flat hospital-based costs, suggesting increased value. This trend potentially reflects improvements in the management of GIB since 2006, such as integration of clinical guidelines recommending restrictive transfusion strategies and early risk assessment with discharge of very-low-risk patients from the ED. The importance of the "Triple Aim" of better health care quality, lower costs, and improved health care outcomes can be seen across different domains in healthcare.[63] This study suggests that improved healthcare delivery for patients with GIB has resulted in captured value for patients, providers, and payers.

There are several limitations to this study. First, NEDS does not include information from the hospital stay for ED patients that were transferred to another hospital, limiting longitudinal follow-up of the clinical management for these patients. However, the majority of patients admitted from EDs with GIB are admitted to the same hospital, which is included in NEDS. Second, we did not have information on patients who may develop GIB after hospitalization for another diagnosis. Third, NEDS only provides encounter-level data and not patient-level data. Therefore, patients with recurrent bleeding who visited the ED more than once would have been counted as separate encounters, leading to an overestimate of GIB incidence. This systematic overestimation should be uniform over the 14-year study period and should not affect the observed trends. Fourth, we used ICD and CPT codes to identify diagnoses and procedures for GIB without individual chart review, which may have introduced some misclassification. This is unlikely to affect specific upper and lower GIB diagnoses as several studies have shown good positive predictive values for these codes.[64-66] However, nearly a third of GIB diagnoses had non-specific ICD codes and could not be categorized into upper or lower GIB.[67] Redistribution of these unspecified cases may affect the relative proportions of upper and lower GIB, especially with the introduction of more specific codes in ICD-10-CM and new diagnostic techniques leading to better identification of bleeding in the small intestine.[68] The shift from ICD-9-CM to ICD-10-CM in 2015 does not primarily explain our findings, given that trends for GIB incidence began before the shift and continued after. Additionally, different secular trends in GIB by age groups also argues against the code change as the main explanation.

Fifth, both NEDS-derived inpatient cost and Medicare DRG-derived cost exclude provider fees and therefore underestimate the true total healthcare cost. Nonetheless, the absence of provider fees is systematic and is unlikely to impact the observed trends. Lastly, case fatality was derived from all-cause death and does not indicate death directly attributable to GIB.

In summary, the overall incidence of acute GIB and UGIB in the U.S. is increasing, especially for young and middle age groups, while the incidence of LGIB is decreasing. Clinical management of GIB appears to reflect updated guideline recommendations, with decreased RBC transfusions and increased patient discharges from the ED. Despite a sicker population presenting with GIB, case fatality rate and inpatient length of stay has decreased with minimal change in healthcare costs. Our findings call for greater awareness and further investigation of underlying causes for the increasing incidence of GIB and UGIB.

Disparities in Access to Endoscopic Evaluation for Patients with Acute Upper Gastrointestinal Bleeding Presenting to the Emergency Department

Introduction

In the United States, acute upper gastrointestinal bleeding (UGIB) has an annual incidence of 72.6 per 100,000 cases, accounts for approximately 230,000 hospitalizations yearly and is associated with an annual in-hospital economic burden greater than $2 billion.[69] The standard of care involves consideration of upper endoscopic evaluation with esophagogastroduodenoscopy (EGD) for patients presenting with acute UGIB with an endoscopist trained in endoscopic hemostasis, possibly within 24 hours of presentation.[5] For Black and Hispanic populations, access to endoscopic evaluation when presenting to the emergency department (ED) with a primary diagnosis of acute UGIB is not clearly described. A previous study found that for hospitalized patients with non-variceal UGIB (NVUGIB), uninsured and Black patients have lower odds of undergoing EGD, Black and Hispanic patients have lower odds of early endoscopic evaluation and Native American patients have the highest odds of in-hospital mortality.[70] Access as measured by services rendered during an ED visit is particularly relevant to Black and Hispanic populations, who are more likely to visit the emergency department (ED) and utilize the ED for routine clinical care.[71]

We aim to assess national disparities in access to endoscopic care among individuals presenting to the ED with UGIB. We hypothesize that historically marginalized racial/ethnic populations will have lower odds of undergoing EGD.

Methods

A retrospective cross-sectional analysis of the 2019 Nationwide Emergency Department Sample (NEDS) was performed. Variables assessed included race/ethnicity, age, sex, hospital region, median income by zip code, insurance, hospital location and teaching status, and Charlson Comorbidity Index (CCI).[44] Univariate, survey adjusted, and population weighted-multivariable logistic regression analyses were performed. The interaction between race/ethnicity, hospital rurality, and hospital teaching status was also assessed with subsequent stratified analyses.

189,547 of 276,740 individuals with a primary diagnosis of UGIB underwent EGD. Most individuals undergoing EGD were White (66.2%), male (54.3%), had Medicare for primary insurance (59.8%), received care in an urban non-teaching hospital (71.0%) and had a Charlson co-morbidity index (CCI) ≥4 (27.0%) (Table 3).

Table 3: Characteristics of Patients with Primary Diagnosis of Upper Gastrointestinal Bleeding Presenting to the Emergency Department.

Characteristics	No Endoscopy	Endoscopy	P value
Total (n)	85557	191183	
Hospital Admission *	38.4	97.5	< 0.001
RBC Transfusion	9.7	36.7	< 0.001
Endoscopic Hemostasis	0.0	34.7	< 0.001
Female	43.9	45.7	< 0.001
Age (year)			< 0.001
< 25	13.4	1.1	
25-45	38.4	15.0	
45-65	31.9	49.6	
> 65	16.3	34.3	
Race/Ethnicity			< 0.001
White	62.4	66.2	
Black	18.1	15.7	
Hispanic	13.5	11.4	
Asian	2.1	3.7	
Native American	0.9	0.4	
Other	3.0	2.7	
Region			< 0.001
Northeast	16.6	18.0	
Midwest	22.8	21.4	

South	36.9	39.7	
West	23.7	20.9	
Teaching Hospital **			< 0.001
Urban Teaching	20.1	21.1	
Urban Non-Teaching	64.4	71.0	
Rural	15.5	7.9	
Income Quartile			< 0.001
$1 - 43,999	32.8	27.8	
$44,000 - 55,999	25.5	24.4	
$56,000 - 73,999	23.5	25.1	
$74,000+	18.1	22.7	
Primary Payer			< 0.001
Medicare	32.8	59.8	
Medicaid	24.1	12.6	
Private Insurance	25.9	19.4	
Self-Pay	13.8	5.6	
No Charge	0.4	0.5	
Other	3.0	2.2	
Charlson Co-morbidity Index			< 0.001
0	55.1	21.1	
1	18.5	21.9	
2	9.0	15.7	
3	6.3	14.3	
4+	11.1	27.0	

*Hospital admission includes patients who were admitted from the ED or placed under observation in the ED

**NEDS does not differentiate teaching status for rural hospitals as only a small number of rural hospitals were teaching

Results

In univariate analysis, patients who were age ≥25 years-old, female, Asian, had a median income by zip code above the lowest quartile, and had a CCI>0 were associated with increased odds of undergoing EGD. Conversely, Black (OR 0.80;CI 0.77-0.84), Hispanic (OR 0.80;CI 0.76-0.84), and Native American (OR 0.41;CI 0.33-0.50) race/ethnicity, and receiving care at a rural hospital (OR 0.49;CI 0.46-0.52) were associated with lower odds of undergoing EGD. On weighted multivariable analysis, Black (OR 0.82;CI 0.78-0.87), Native American (OR 0.60;CI 0.47-0.76), Medicaid (OR 0.74;CI 0.70-0.79) and patients receiving care in rural/non-metropolitan hospitals (OR 0.51;CI 0.48-0.55) had lower odds of undergoing EGD. In comparison, increasing age and median income by zip code, Asian race (OR 1.7;CI 1.49-1.94), and receiving care in urban teaching-hospitals (OR 1.06;CI 1.01-1.11) were independent predictors for undergoing EGD (Table 4).

Table 4: Weighted Univariate and Multivariate Logistic Regression Analysis for Upper Endoscopic Evaluation in Patients with Upper Gastrointestinal Bleeding

	Univariate		**Multivariate**	
Weighted Population = 276,744				
Number of Encounters = 64,330				
	Odds Ratio	95% Confidence Interval	Odds Ratio	95% Confidence Interval
Age				
<25 years old	Reference	Reference	Reference	Reference
25-50 years old	4.75***	4.27-5.28	3.75***	3.37-4.17
50 to 75 years old	18.86***	16.99-20.93	9.8***	8.79-10.92
>75 years old	25.56***	22.96-28.46	10.9***	9.68-12.27

Gender				
Male	Reference	Reference	Reference	Reference
Female	1.08***	1.04-1.11	0.94**	0.91-0.98
Region				
Northeast	Reference	Reference	Reference	Reference
Midwest	0.87***	0.82-0.91	1.03	0.97-1.10
South	0.99	0.95-1.05	1.2***	1.13-1.27
West	0.81***	0.77-0.86	0.86***	0.81-0.92
Zip Income Quartile				
$1-47,999	Reference	Reference	Reference	Reference
$48,000 - 60,999	1.13***	1.08-1.18	1.06*	1.00-1.12
$61,000 - 81,999	1.26***	1.20-1.32	1.13***	1.07-1.20
$82,000+	1.48***	1.40-1.55	1.22***	1.14-1.29
Primary Payer				
Medicare	Reference	Reference	Reference	Reference
Medicaid	0.29***	0.27-0.30	0.72***	0.68-0.77
Private Insurance	0.41***	0.39-0.43	0.99	0.93-1.05
Self-Pay	0.22***	0.21-0.24	0.69***	0.63-0.74
No Charge	0.61***	0.47-0.78	1.42*	1.07-1.89
Other‡	0.4***	0.36-0.45	0.82**	0.73-0.93
Charlson Co-morbidity Index				
0	Reference	Reference	Reference	Reference
1	3.09***	2.95-3.25	2.14***	2.03-2.25
2	4.55***	4.28-4.84	2.54***	2.38-2.71

3	5.99***	5.58-6.42	3.56***	3.30-3.83
≥4	6.40***	6.02-6.72	3.32***	3.12-3.52
Urban vs Rural and Teaching Status				
Urban (>50,000) and Non-Teaching Hospital	Reference	Reference	Reference	Reference
Urban (>50,000) and Teaching Hospital	1.05*	1.01-1.10	1.06*	1.01-1.11
Rural (<50,000) and Non-Metropolitan Hospital	0.49***	0.46-0.52	0.51***	0.48-0.55
Race				
White	Reference	Reference	Reference	Reference
Black	0.81***	0.78-0.85	0.84***	0.79-0.89
Hispanic	0.79***	0.75-0.84	0.98	0.92-1.04
Asian	1.66***	1.48-1.87	1.7***	1.49-1.94
Native American	0.4***	0.33-0.50	0.6***	0.47-0.76
Other	0.86**	0.78-0.95	1.01	0.90-1.14

* p<0.05, ** p<0.01, *** p<0.001

‡ Includes Worker's Compensation, CHAMPUS, CHAMPVA, Title V, and other government programs

On stratified analysis Black race was associated with lower odds of undergoing EGD in urban-teaching hospitals (OR 0.79;CI 0.75-0.85) and urban non-teaching hospitals (OR 0.87;CI 0.77-0.98), while Hispanic ethnicity (OR 0.49;CI 0.34-0.72) or Native American race (OR 0.46;CI 0.28-0.76) were associated with lower odds of undergoing EGD in rural hospitals. (Table 5)

Table 5: Stratified multivariate analysis of upper endoscopy for patients with UGIB by Urban / Rural and Teaching Hospital Status

Race		**Urban Non-Teaching Hospital**		**Urban Teaching Hospital**		**Rural Hospital**	
	N	**OR**	**95% CI**	**OR**	**95% CI**	**OR**	**95% CI**
White	179949	Ref	Ref	Ref	Ref	Ref	Ref
Black	45452	0.87*	0.77-0.98	0.79***	0.74-0.84	1.02	0.83-1.25
Hispanic	33252	1.11	0.97-1.27	0.97	0.90-1.04	0.49***	0.34-0.72
Asian	8866	1.40*	1.07-1.84	1.71***	1.47-1.99	2.92**	1.40-6.11
Native American	1491	0.53*	0.29-0.935	0.85	0.60-1.20	0.46**	0.28-0.76
Other	7734	1.00	0.78-1.27	1.04	0.91-1.20	0.77	0.37-1.61

Multivariate weighted logistic regression adjusted for age (per 5-year increase), sex, hospital region, income quartile, CCI score, and year

Discussion

In this study we found that Black and Native American patients had lower odds of undergoing EGD for UGIB compared to White patients. When stratified by context, Black patients had lower odds of undergoing EGD for UGIB compared to White patients in urban-teaching and non-teaching hospitals. In addition, Hispanic and Native American patients had lower odds of undergoing EGD for UGIB compared to White patients in rural hospitals. Conversely, Asian race and receiving care in an urban teaching-hospital were independent predictors for undergoing EGD for UGIB. Our analysis found that Black and Native American populations have lower odds of undergoing EGD for UGIB, which is consistent with prior research showing that Black patients have lower odds of undergoing EGD and early endoscopic evaluation.[70] Similar to prior studies, we also found that Asian populations have higher odds of undergoing EGD for UGIB.[70] While socioeconomic factors and co-morbidity burden disparately effect historically marginalized populations, the disparity in access persists after our adjustment for CCI, primary insurance payer, and zip code income quartile. Explanations for this disparity could include individual barriers (time-off, lost wages and ability to secure affordable childcare) and structural barriers (healthcare fluency, mistrust of the healthcare system, implicit bias, and structural racism). Healthcare fluency refers to the conglomerate of general literacy, scientific knowledge, cultural perception as well as trust and ability to self-advocate in the healthcare system.[72] Healthcare fluency may thus effect the perceived understanding regarding urgency of an acute UGIB and the recommendation to undergo potential EGD.

Implicit bias can lead to varying treatment recommendations across race/ethnicity.[73] Importantly, these biases may be exacerbated under periods of high stress,[74] such as caring for patients with an acute UGIB. Implicit bias may subsequently contribute towards Black, Hispanic, and Native American populations having worse quality healthcare measures.[75] While our findings do not explicitly link these individual and structural barriers to disparities in access, structural, interpersonal, and internalized racism is a modifiable risk factor that may be addressed through anti-racism policies[76] in order to combat ongoing inequities in care.

Our study showed that receiving care in a rural/non-metropolitan hospital was independently associated with lower odds of undergoing EGD for UGIB. Rural hospitals are being increasingly classified with safety-net status[77] and safety-net hospitals disproportionately care for low-income, uninsured/underinsured, and historically marginalized racial/ethnic populations. Thus, this finding may be in the context of both safety-net hospitals often being under-resourced and rural hospitals also facing geographic isolation with limited access to specialty care. Study findings also demonstrate increased odds of undergoing EGD in urban-teaching hospitals, which may be due to increased access to sub-specialty care.

Furthermore, stratified analysis assessing the interaction between race/ethnicity, hospital rurality, and hospital teaching status demonstrates that Hispanic ethnicity and Native American race was associated with lower odds of undergoing EGD in rural hospitals. In the US, half of the agricultural workforce, often located in rural regions, is comprised of Hispanic individuals[78] and 34% of agricultural laborers are undocumented.[78] Additionally, rural EDs are experiencing greater utilization rates compared to urban EDs and are increasing their care of Medicaid, and uninsured patient populations.[77] Unfortunately, rural Hispanic populations have the lowest rates of insurance coverage and the highest rates of not having a healthcare provider.[79] Hispanic individuals also have lower endoscopic therapy rates[70] and the highest rebleeding rates.[80] Our study findings could be explained, in part, by structural barriers impacting the patient-provider dynamic and subsequent provision of care, which may also be compounded by documentation and insurance status.

Race/ethnicity has been found to be an independent risk factor for differences in specialty healthcare delivery, which can contribute to inequities in outcomes.[81] Our study found that historically marginalized patient populations including Black, Native American and individuals receiving care in a rural hospital are independently associated with lower odds of undergoing EGD for UGIB. Additionally, Hispanic, and Native American patients have lower odds of undergoing EGD in rural hospitals. While individual and structural barriers may contribute to this disparity, further studies should assess barriers on the continuum of care for UGIB such as appropriate access to endoscopic care and equitable healthcare delivery, including policies governing access to procedural resources.

Cost Minimization Analysis of Applying Risk Stratification to Patients Presenting with Acute Upper Gastrointestinal Bleeding

Introduction

Approximately 300,000 patients present to the emergency department (ED) in the U.S. with upper gastrointestinal bleeding (UGIB) annually. National guidelines recommend using risk assessment scores, such as the Glasgow-Blatchford Score (GBS), to discharge very low risk patients, who have a very low risk of requiring red blood cell transfusion, hemostatic intervention, or death.[82,83] In a prospective non-randomized multicenter study studying the GBS, 29% of patients were discharged, with none of them experiencing any adverse events (death, interventions or readmission).[84]

Despite this evidence, uptake has been low, with approximately only 30% of all physicians ever using an upper gastrointestinal bleeding risk score in a national survey of emergency physicians, internists, and gastroenterologists practicing in hospitals affiliated with an ACGME-accredited gastroenterology fellowship.[20] Underutilization of risk stratification scores has been attributed to lack of knowledge, unbelief in the value of using these scores, and diffusion of responsibility for risk classification among ER physicians, gastroenterologists, and nurses.[21]

Previous studies have shown that integrating clinical decision support within electronic health records can make implementation of new practices more salient and overcome clinician inertia in adapting unfamiliar practices. Furthermore, these approaches may be more easily scaled, facilitating the use of risk adjustment tools such as GBS, or even more powerful machine learning tools, into care for patients with UGIB.

Quantifying the cost implications of consistently using risk stratification for patients with UGIB is important. Implementing and maintaining clinical decision support applications comes at a cost to each hospital system—a prior study has estimated the cost of a similar application at $217,138.[85] Nevertheless, these costs are offset by averted hospitalizations, which result in lower inpatient spending, fewer days of work lost, and less informal care.

This study evaluates the potential healthcare and societal cost implications of widespread implementation of GBS or a machine learning model to guide triage of patients with UGIB using a cost minimization analysis. We account for the cost of implementation and maintenance of clinical decision support applications, as well as cost-savings related to reduced inpatient utilization, reductions in lost employment, and lower cost of informal care. These analyses form a robust framework that can provide a comprehensive estimate of the potential cost savings to society with the consistent use of risk stratification tools in UGIB.

Methods

Model Structure

A hypothetical cohort of patients with UGIB was modeled using a Markov chain model that followed the cohort from presentation to the ED through either inpatient stay to 28 days post-discharge, or 28 days after discharge from the ED. This timeframe was chosen because generally, post-discharge diagnostic evaluation and assessment with endoscopy, if necessary, is completed within 30 days of the patient encounter. Moreover, the patients modeled (very low risk), are not expected to have complications past this timeframe.

Two triage strategies were compared to usual care: using GBS=0 or a previously validated machine learning model applied for all patients presenting with UGIB with comparable sensitivity to identify very low risk patients. The use of GBS=0 applied to every patient with UGIB reflects the multicenter prospective trial where 71% of patients with UGIB were admitted compared to the previous rate of 96%. The use of machine learning estimated proportion of admitted patients at 65% using the absolute increase in sensitivity of the machine learning model (14%) compared to GBS at the matched 100% sensitivity threshold in an external validation study. Since this was not directly taken from the original trial, the proportion of admitted patients was varied along a normal distribution with standard deviation of 5%. For each strategy, patients are either admitted to the hospital or discharged from the ED for outpatient care.

Triage Strategies

For each scenario, we compared usual care with application of the Glasgow Blatchford Score and a previously published machine learning model at the very high sensitivity threshold identify very low risk patients.

Model Inputs

Rates of hospital admission from the ED, in-patient mortality, discharge from the hospital, discharge from the ED, post-ED discharge mortality, post-ED readmission, and outpatient endoscopy under each triage strategy were derived from a prospective multicenter trial.[84] Proportion of 30-day readmissions were derived from a retrospective study using the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project 2014 Nationwide Readmission Database for hospitalized patients with non-variceal UGIB, and the proportion adjusted for the projected decrease in hospital admissions with the applied triage tool.[86] The proportion of patients discharged from the hospital without endoscopic evaluation, and thus potentially needing outpatient endoscopic evaluation, was derived from the 2019 National Emergency Department Sample. Of patients who require outpatient endoscopy, the proportion of patients who go to an Ambulatory Surgery Center versus a Hospital Outpatient Surgery Center were derived from a report published by the Anthem Public Policy Institute in 2020. For the patient perspective to quantify indirect costs, the rate of patients requiring informal care with lost employment were derived from the largest prospective micro-costing study of patients with AUGIB from the TRIGGER pragmatic cluster randomized trial of restrictive versus liberal RBC transfusion strategies for AUGIB.[87]

Costs from Payer Perspective

Our analysis adopts a U.S. health care sector perspective and evaluates direct costs, including validation, implementation and maintenance of risk assessment scores. Inpatient costs and outpatient follow-up costs (clinic visit, outpatient laboratory values, endoscopy) were estimated from Medicare reimbursement tables (Table 6).

Table 6: Key Input Parameters with references and Costs based on Medicare reimbursement cost

Parameter	Base Case Value	Source	Distribution
Admission Rate from ED			
Usual Care	0.96	Stanley et al. 2009	
GBS	0.71	Stanley et al. 2009	
Machine Learning Model	0.65	Shung et al. 2020	Normal
Proportion of Admitted Patients who Die	0.04	Stanley et al. 2009	
Patients Discharged from Inpatient Stay Readmitted to the ED			
Usual Care	0.13	Abougergi et al. 2018	
GBS	0.19	Abougergi et al. 2018	
Machine Learning Model	0.21	Abougergi et al. 2018	
Patients Discharged from Inpatient Stay without endoscopy performed inpatient	0.309	National Emergency Department Sample 2019	
Patients Discharged from Inpatient Stay who Return for Outpatient Endoscopy	0.4	Stanley et al. 2009	
Ambulatory Surgery Center	0.52	Anthem Public Policy Institute 2020	
Hospital Outpatient Surgery Center	0.48	Anthem Public Policy Institute 2020	
Discharge Rate from ED			
Usual Care	0.04	Stanley et al. 2009	
GBS	0.29	Stanley et al. 2009	
Machine Learning Model	0.35	Shung et al. 2020	
Proportion of Discharged Patients who Die	0	Stanley et al. 2009	
Proportion of Discharged Patients who Return for Outpatient Endoscopy	0.4	Stanley et al. 2009	
Ambulatory Surgery Center	0.52	Anthem Public Policy Institute 2020	
Hospital Outpatient Surgery Center	0.48	Anthem Public Policy Institute 2020	
Patients Discharged from Inpatient Stay who Require Informal Care	0.344	Campbell et al. 2015	
Cost			
Year 1 Validation and Implementation	$2,867.29	Sendak et al. 2017	
Years 2-10 Maintenance	$519.61	Sendak et al. 2017	

We chose to use Medicare reimbursement cost because Medicare pays less than other insurers and is a conservative estimate of cost savings when applying either the GBS or the machine learning model. For inpatients, we used a weighted cost derived from Diagnosis Related Groups for gastrointestinal bleeding: 377, 378, and 379 with proportions of patients with each DRG derived from the 2019 NEDS and national average Medicare Payment Amount from 2021 Medicare tables. Provider fees were calculated based on length of stay estimates for DRG 378 and 379 with Hospitalist and Gastroenterology services rendered, while for DRG 377 we assumed 60% required ICU care for half of their stay and 40% did not. Provider fees for inpatient upper endoscopy with conscious sedation fee were estimated with proportions from 2019 NEDS for diagnostic EGD only, EGD with biopsies, EGD with hemostasis, and EGD with band ligation. For outpatients, we assumed patients discharged from the hospital would require one outpatient clinic visit and one laboratory draw comprising of complete blood count, complete metabolic panel, and prothrombin time. We assumed that patients discharged from the ED may require double the visit and laboratory testing. Costs for outpatient endoscopy with conscious sedation fee in ASC versus HOSC were estimated from Medicare reimbursement tables and rates assumed to be 30% diagnostic and 70% with biopsy. Total initial validation and implementation costs ($217,519) in an EHR for year 1 was estimated from an analytics application for chronic kidney disease to identify high risk patients requiring a nephrology referral.[85] Maintenance costs were estimated from the same study and quantified based on the need for query development to extract variables, data exploration and data pipeline costs (total $39,350 each year) for years 2 to 10 per hospital emergency department, which were applied on a per-patient basis. The per-patient basis was calculated using the 2019 National Emergency Department Sample. Additional details are provided below.

Inpatient Costs:

<u>DRG Costs:</u>

For DRG 379, GI Hemorrhage without CC/MCC the 2021 national average Medicare payment was $4,056.92. 2019 NEDS estimated 6.6% of patients with hospitalization for UGIB.

DRG 378, GI Hemorrhage with CC the 2021 national average Medicare payment was $6,421.23. 2019 NEDS estimated 47.6% of patients with hospitalization for UGIB.

DRG 377, GI Hemorrhage with MCC the 2021 national average Medicare payment was $12,326.26. 2019 NEDS estimated 25% of patients with hospitalization for UGIB.

The weighted cost was calculated by adding the proportion of each DRG multiplied by the national average Medicare payment for DRGs 377, 378, and 379.

Inpatient Provider Fees

Provider fees were estimated by associated mean length of stay for each DRG from the 2019 NEDS. For DRG 379, the length of stay was 2.23 (CI 2.18-2.28). On Day 1 Hospitalist and Gastroenterologist would both bill 99222 for Initial Hospital Care; Day 2 the Hospitalist and Gastroenterologist would bill 99231; Day 3 the Hospitalist would bill 99238 Hospital Discharge day.

For DRG 378 the length of stay was 3.37 (CI 3.34-3.40). On Day 1 Hospitalist and Gastroenterologist would both bill 99222 for Initial Hospital Care; Day 2 the Hospitalist and Gastroenterologist would bill 99232; Day 3 the Hospitalist and Gastroenterologist would bill 99231; Day 4 the Hospitalist would bill 99238 Hospital Discharge Day and Gastroenterologist would bill 99231.

For DRG 379 the length of stay was 5.72 (CI 5.63-5.81). We assumed that 60% would require ICU care for 50% of their hospitalization (3 days out of 6), and 40% would not.

For DRG 379 requiring ICU care: on Day 1 Hospitalist or ICU would bill 99291 and Gastroenterologist would bill 99223 for Initial Hospital Care; Day 2 Hospitalist or ICU would bill 99291 and Gastroenterologist would bill 99232; Day 3 Hospitalist or ICU would bill 99291 and Gastroenterologist would bill 99232; Day 4 the Hospitalist and Gastroenterologist would bill 99232; Day 5 the Hospitalist and Gastroenterologist would bill 99231; Day 6 the Hospitalist would bill 99239 Hospital Discharge Day.

For DRG 379 not requiring ICU care: on Day 1 Hospitalist and Gastroenterologist would both bill 99223 for Initial Hospital Care; Day 2 the Hospitalist would bill 99233 and Gastroenterologist would bill 99232; Day 3 the Hospitalist would bill 99233 and Gastroenterologist would bill 99232; Day 4 the Hospitalist and Gastroenterologist would bill 99232; Day 5 the Hospitalist and Gastroenterologist would bill 99231; Day 6 the Hospitalist would bill 99239 Hospital Discharge Day.

CPT	**CPT Code Description**	**2021 National Medicare Rate**
Critical Care Services-Inpatient Only		
99291	Critical Care First Hour	$220.87
99292	Critical Care Additional 30 Min	$110.96
Hospital Care, Inpatient		
99221	Initial Hospital Care	$101.19
99222	Initial Hospital Care	$136.08
99223	Initial Hospital Care	$200.29
Subsequent Hospital Care, Inpatient		
99231	Subsequent Hospital Care	$38.38
99232	Subsequent Hospital Care	$71.88
99233	Subsequent Hospital Care	$103.28
Hospital Discharge, Inpatient		
99238	Hospital Discharge Day	$72.23
99239	Hospital Discharge Day	$106.42
Admission & Discharge on Same Day from Hospital Observation Care		
99234	Observation/Hospitalization Same Date	$131.55

99235	Observation/Hospitalization Same Date	$167.14
99236	Observation/Hospitalization Same Date	$214.59
99217	Observation Care Discharge	$72.23
Outpatient GI Consult		
99204	Office New	$137.48

Inpatient EGD Provider Fees

UGIB patients who underwent diagnostic EGD only (19.8%), EGD with biopsy (35.1%), EGD with hemostasis (22.5%), and EGD with ligation (1.9%) were derived from 2019 NEDS and multiplied by the 2021 National Average Medicare Provider Fee to obtain a weighted estimate. We added the conscious sedation fee to all EGDs since the estimate was easily obtained and anesthesia fees were not clearly delineated in Medicare.

Inpatient Endoscopy Fee	**Description**	**Provider Fee only**
45235	EGD only	$124.57
43239	EGD with biopsy	$140.27
43255	EGD and hemostasis	$203.43
43244	EGD and ligation	$248.44
Conscious Sedation Fee		
99151	Conscious Sedation	$25.47

Observation Provider Fees

From 2019 NEDS 4.4% of patients with UGIB were under observation, and we assumed billing of 99235 (Observation Same Date), 99217 (Observation Care Discharge), and 99204 (GI consult while observation).

Outpatient Endoscopy Costs

We assumed that of all patients presenting for outpatient endoscopy, 30% had diagnostic endoscopy only and 70% had endoscopy with biopsy, presumably to test for H. pylori.

CPT Code	**ASC**	**HOSC**
45235	$533.00	$933.00
43239	$549.00	$949.00

Per-Patient Cost of Implementation and Maintenance

In the 2019 NEDS, there were 920 unique hospital EDs in the sample for UGIB, with 69,671 patient encounters (unweighted). Based on the cost of implementation of $217,138 multiplied by 920 EDs was divided by the 69,671 unweighted patient encounters to give a cost per patient of $2,867.29. Likewise, maintenance cost per patient was calculated to be $519.61.

Costs from Patient Perspective

Our analysis also performs a secondary model incorporating indirect costs faced by patients after inpatient hospitalization, namely lost employment and costs of informal care. The mean hours of unpaid informal care were 69.17 hours with standard deviation of 10.3 hours, and lost working hours 125.21 hours with standard deviation of 7.8 hours. Per patient cost was estimated using the estimated average annual wage by the Economic Policy Institute and assuming 52 40-hour work weeks.

Main Outcomes

Outcomes were per patient costs and per patient length of stay over the first 10 years of implementation.

Sensitivity Analyses

We performed several scenario analyses, including patient-borne indirect costs and inpatient costs estimated from the 2019 NEDS rather than Medicare tables. We performed probabilistic sensitivity analysis sampling model inputs 1,000 times from uncertainty distributions, and tested model stability by modeling our base-case scenario with 10,000 simulations. We performed model testing using TreeAge Pro Healthcare Version 2021 R2.0 release.

Results

Base Case Analysis

With the GBS threshold of 0, projected cumulative savings at year 2 after implementation are $346.92 and at year 3 $1694.52 per patient when the score is applied compared to usual care (Figure 8). Projected savings using a machine learning model at a matched sensitivity threshold are $1248.27 at year 2 and $3048.97 at year 3 per patient.

Scenario Analyses

<u>Base Case with Indirect Costs</u>

Incorporating indirect patient-facing costs of informal care and lost employment, with the GBS threshold of 0, projected cumulative savings at year 2 after implementation are $785.52 and at year 3 $2133.12 per patient when the score is applied compared to usual care (Figure 8). Projected savings using a machine learning model at a matched sensitivity threshold are $1786.16 at year 2 and $3586.86 at year 3 per patient.

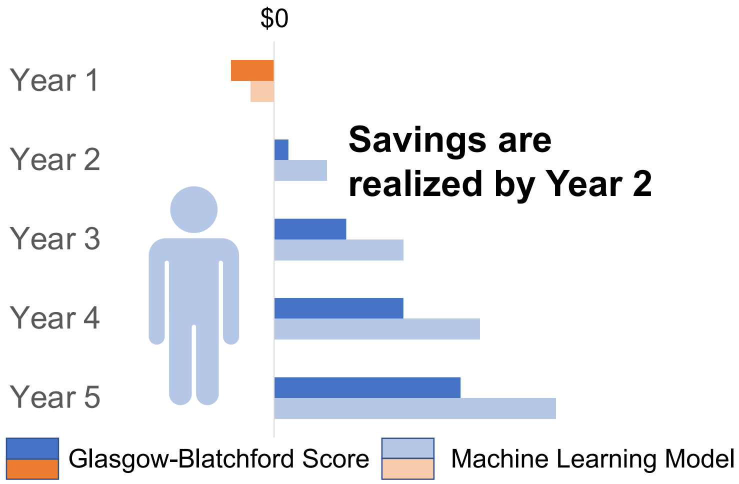


Figure 8: Projected cumulative savings on a per-patient basis in the first 5 years for a single healthcare system.

When applied across the United States, a projected 1 billion dollars could be saved in cumulative savings by Year 5 with either GBS or a machine learning model. (Figure 9)

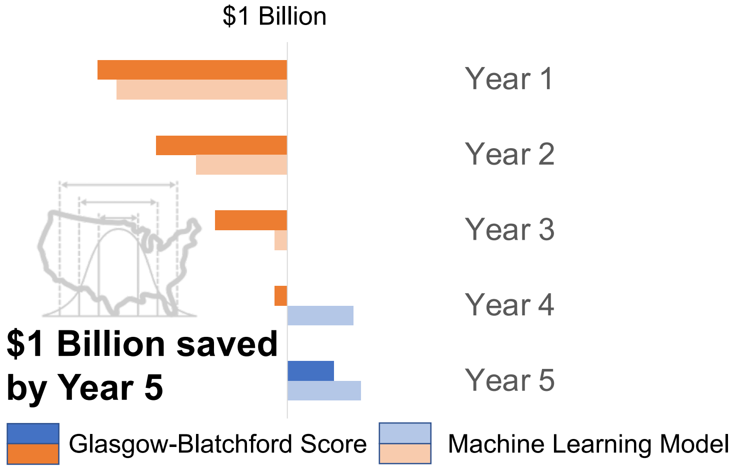


Figure 9: Projected cumulative savings across the United States using the National Emergency Department Sample 2019 weighted estimate of patients presenting with a primary diagnosis of upper gastrointestinal bleeding and number of healthcare systems.

NEDS Costs

Using 2019 NEDS costs instead of Medicare costs for the inpatient associated costs, with the GBS threshold of 0, projected cumulative savings at year 2 after implementation are $614.02 and at year 3 $1961.62 per patient when the score is applied compared to usual care (Figure 8). Projected savings using a machine learning model at a matched sensitivity threshold are $1574.96 at year 2 and $3375.66 at year 3 per patient.

NEDS Costs with Indirect Costs

Using 2019 NEDS costs instead of Medicare costs for the inpatient associated costs, with the GBS threshold of 0, projected cumulative savings at year 2 after implementation are $876.32 and at year 3 $2223.92 per patient when the score is applied compared to usual care (Figure 8). Projected savings using a machine learning model at a matched sensitivity threshold are positive at year 1 with savings of $118.92 per patient, $1919.62 at year 2 and $3720.32 at year 3 per patient.

Discussion

Full implementation of GBS or machine learning risk assessment model for patients presenting with UGIB is projected to be cost-saving for the U.S. healthcare sector by year 2, with potential nationwide savings of hundreds of millions of dollars annually by year 3, and 1 billion dollars by year 5. Payers should consider developing novel payment structures to incorporate reimbursement to healthcare systems for the use of clinical decision support tools in patients with UGIB.

Under scenario analyses, the incorporation of indirect patient-facing costs and use of 2019 NEDS estimate for inpatient costs led to a positive cost savings within year 1 with implementation of a machine learning model for UGIB.

Development and Validation of Deep Learning and Machine Learning Approaches versus Clinical Risk Scores using Electronic Health Records

Introduction

Gastrointestinal bleeding (GIB) is the most common gastrointestinal diagnosis requiring hospitalization in the U.S., and accounts for 2.2 million hospital days and inpatient charges of 19.2 billion dollars[69] Traditionally acute GIB has been classified by the suspected or confirmed anatomical location of the bleeding, either upper GIB or lower GIB. Guidelines recommend risk stratification using clinical risk scores for both upper and lower gastrointestinal bleeding to identify very-low-risk patients, defined as patients who do not require red-blood-cell transfusion or intervention to stop bleeding, and who not die.[5-7,9,88] Once accurately identified as very-low-risk, such patients can then be considered for discharge from the emergency room with outpatient management, thereby reducing costs without risk of harm to the patient. Existing clinical risk scores for upper and lower GIB are used uncommonly in clinical practice. Barriers to use include cumbersome data entry, an uncertain distinction between upper and lower sources of GIB at presentation in many patients, as well as relatively poor performance of many risk scores.

The location of the bleeding source is not always clear at presentation: a patient presenting with melena may have a bleeding lesion in the small intestine or proximal colon, while hematochezia can be the presenting symptom for up to ~15% of patients with upper GIB.[89] Therefore, risk assessment tools that are designed for patients with symptoms of GIB rather than site of bleeding would be more clinically useful and practical for assessment of risk in the emergency room.[90] The Glasgow Blatchford Score has been validated on multiple patient cohorts with upper GIB and is able to identify very-low-risk patients with high sensitivity (low false negative rate) allowing discharge of these patients from the emergency room without hospitalization.[12,84] The Oakland score is a clinical risk score recently developed for lower GIB and has shown good performance for identifying very-low-risk patients when validated in a U.S. electronic health record (EHR) database.[15,16]

Machine learning, a field of study that gives computers the ability to learn without being explicitly programmed, can be used to construct models that perform better than clinical risk scores for GIB.[13] Deep learning models, which use layers of computational units to create a complex function from simpler functions, appear to have better overall performance than machine learning models.[13] Furthermore, the EHR stores a large quantity of clinical data that can be used to automatically calculate risk assessments. Such a model ideally could be applied to all patients with GIB, preventing the need for immediate differentiation[91,92] between upper and lower GIB. We previously showed a machine learning model performed better than standard risk stratification tools in identifying very-low-risk patients with upper GIB, but this analysis used only 24 variables obtained by manual data collection.[14] We propose a symptom-based deep learning risk score derived from EHR data that bases initial assessment on presenting symptoms rather than presumed anatomical location. This score is designed to be deployed automatically through the EHR by extracting data to automatically calculate scores that can be then made available to providers in real-time, soon enough after presentation to be of clinical utility in making important decisions regarding care. We compare the best performing machine learning model from previous work to the deep learning model and two guideline-recommended clinical risk scores, the Glasgow Blatchford Score and the Oakland Score.

Methods

Data Source

A cohort of patients presenting with overt GIB, defined as hematemesis, melena, or hematochezia, from July 2014 to December 2017 to the emergency rooms of the two campuses of Yale New Haven Hospital in New Haven, CT was used for development and temporal validation of the machine learning model. A separate cohort of patients presenting with overt GIB to the emergency room of a separate hospital (Bridgeport Hospital, Bridgeport, CT) from July 2014 to December 2017 was used for external validation. EHR phenotyping for overt GIB using a Boolean decision rule integrating GIB-specific triage terms (structured datafields entered by the emergency room triage nurse) and GIB-specific review-of-systems fields was utilized to identify a total of 5,720 patients at these hospitals from 2014 to 2017 who met the criteria of having complete data available at 4 hours into the emergency department encounter. Manual validation with chart review was performed for this sample to evaluate for overt GIB in the emergency room rather than an episode of over GIB during the hospital stay, resulting in the final patient cohorts (total of 3,572).

Study Design

A cohort of 2,004 patient encounters from January 2014 to December 2016 was used as the development set to train and tune (internally validate) the model. In order to evaluate model performance in the same centers, a cohort of 719 patient encounters from January 2017 to January 2018 was used to test the model via temporal validation. The model was then externally validated on a separate cohort of 849 patient encounters from 2014 to 2017 at a geographically separate hospital.

Input Variables

The model used 151 structured datafields in the EHR (Epic, Verona, WI) that were available within 4 hours of presentation to the emergency room. There were 1,931 unique laboratory tests that were sorted by the base name to 701, 3,654 unique medication classes sorted by generic code to 1,198, 3,283 unique ICD-10 codes in Medical History, and 9 specific data elements including vitals and nursing assessments. Feature selection to the final set was a combination of expert-driven selection of relevant laboratory values, vital signs, and provider assessment plus a selection of features from the medical history and medication tables. The features from medical history and medication tables were selected using a pre-specified cutoff of 95% to exclude very rare variables from the final feature set. (Table 7) The final dataset includes 20 base laboratory tests, 70 medication classes, 6 variables from 3 vital signs, 3 variables from nursing assessments, age, gender, and 50 variables from ICD-10 codes listed in the medical history.

Table 7: Input Variables Included in the Models

Demographic Information	Gender, Age at Encounter
Vital Signs and Nursing Assessments	Systolic Blood Pressure Diastolic Blood Pressure Pulse Pulse Oximetry (oxygen saturation) Respiratory Rate Temperature Glasgow Coma Scale – Eye, Motor, anc Assessments
Medication Classes	Alpha-2-receptor antagonist antidepress Alpha-Beta adrenergic blocking agents Analgesic antipyretics salicylates Analgesic antipyretics non salicylate Antacids Anti anxiety benzodiazepines Anticholinergics (orally inhaled long acti Anticoagulant Coumadin Anticonvulsant benzodiazepine Anticonvulsants Antidiarrheal microorganism agents Antiemetic antivertigo agents Antihistamines 1st generation Antihistamines 2nd generation Antihyperglycemic (biguanides) Antihyperglycemic (insulin release stimu Antihyperlipidemic (HMG CoA reductase statins) Antihypertensives (ACE inhibitors) Antihypertensives (angiotensin receptor Antipsychotic (atypical dopamine seroto antagonist) Benign prostatic hypertrophy micturition Beta adrenergic agents (inhaled short a Beta adrenergic and anticholinergic con Beta adrenergic and glucocorticoid com Beta adrenergic blocking agents Blood sugar diagnostics Calcium channel blocking agents Calcium replacement Direct Factor Xa inhibitors Durable medical equipment (miscellane Electrolyte depleters Folic acid preparations Glucocorticoids Histamine H2 receptor inhibitors Xanthine oxidase inhibitors Insulins Iron replacement Laxatives and cathartics

Outcome Variable

The primary outcome was a composite of red-blood-cell transfusion, hemostatic intervention (endoscopic, interventional radiologic, or surgical), and 30-day mortality.

Data Processing

Continuous variables that include laboratory tests, vital signs, and age are extracted with complete cases. Log transformation prior to standardization was performed on the variables of oxygen saturation, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total bilirubin, blood urea nitrogen, creatinine, international normalized ratio, and white blood cell count. Standardization alone was performed on age at encounter, diastolic blood pressure, systolic blood pressure, heart rate, respiratory rate, temperature, albumin, chloride, bicarbonate, hematocrit, hemoglobin, potassium, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, platelet count, red blood cell distribution width, and sodium. Categorical variables of nursing assessment in the Emergency Department include Glasgow Coma Score for eye response, motor activity, and verbal response.

Missing Values

Medication classes and medical history diagnoses are treated as indicator variables that code missing values as 0 and presence as 1. Otherwise, we limited our dataset to patients with complete data.

Model Background and Comparisons

Initially we assessed the performance of the neural network model compared to four machine learning models: gradient boosted tree, regularized regression with the elastic net penalty, and random forest. The neural network model had equivalent overall performance with the gradient boosted tree model and was used for all analyses to predict the composite outcome for patients presenting with acute GIB (Table 8).

Table 8: Performance characteristics of different machine learning and deep learning models.

	Neural Network Model AUC (95% CI)	**Gradient Boosted Decision Tree Model AUC (95% CI)**	LASSO AUC (95% CI)	Random Forest (1000 trees) AUC (95% CI)
External Validation (N = 849)	**0.92 (0.90-0.94)**	**0.92 (0.90-0.94)**	0.91 (0.89-0.93)	0.91 (0.89-0.93)

These models were developed and tuned using the glmnet and randomForest packages in R (R Foundation for Statistical Computing, Vienna, Austria). The gradient boosted tree model assembles a collection of decision trees (tree ensemble model) by adding trees that minimize prediction error measured by a gradient at each training step. Specifically, the model used also has additional options of regularization, shrinkage, and subsampling to prevent overfitting and speed up computation time. The XGBoost package in R and Python (Python Software Foundation) was used to develop and tune the final algorithm. Hyperparameters were tuned using a grid-search approach, with final parameters presented below:

Parameters for XGBoost to Develop Final Model: Learning rate (eta) = 0.01, Minimum split loss (gamma) = 0.013, Maximum tree depth (max_depth) = 18, Minimum sum of instance weight in a child (min_child_weight) = 0, Subsample ratio of training instances (subsample) = 0.75, Subsample ratio of columns when constructing each tree (colsample_bytree) = 0.75, L1 regularization (alpha) = 0.6, L2 regularization (lambda) = 0.01

For the deep learning neural network model, we trained a 5-layer feedforward neural network with ReLU activation functions attached to a binary classifier to predict outcome. The weights were initialized using the Kaiming initialization and trained using stochastic weight averaging with lookahead optimizer with early stopping. A hyperparameter search for learning rate, L1/L2 regularization, dropout rate, and AdamW optimizer was performed using a Tree-structured Parzen estimator using the Optuna framework.

Statistical Analysis

Primary analysis of model performance was assessment of area under receiver operating characteristic curve (AUC), with predefined goal of AUC>0.90. The AUC on internal validation was reported with tenfold cross validation, and the AUCs on temporal validation and external validation were compared using the nonparametric DeLong test. Secondary analysis was specificity when sensitivity is 99% or higher. Specificity identifies the proportion of patients who will not die or need transfusion or intervention (i.e., very-low-risk patients) who are correctly identified by the risk score as being at very-low-risk; the higher the specificity the more patients who can be discharged from the emergency room with very low risk. Comparison of specificities was performed at the 99% sensitivity threshold for external validation (versus GBS=0, Oakland=8). Specificities were compared using McNemar's test. Chi-squared and T-tests were used to compare the variables and characteristics of training and test patient cohorts. Calibration was measured by Brier score for the deep learning model, the machine learning model, the GBS, and the Oakland score.

Results

Selected baseline characteristics of the cohorts used for development, test set, and external validation of the model are shown in Table 9.

Table 9: Selected characteristics of datasets used in development and validation of machine learning model

	Development Dataset[1]	Temporal Dataset[1]	External Validation Dataset[1]
Number of Patients	2004	719	849
Age	65 (52, 79)	67 (53, 79)	67 (54, 80)
Sex			
Female	935 (47%)	322 (44%)	379 (44%)
Male	1,069 (53%)	402 (56%)	473 (56%)
Race			
White or Caucasian	1,332 (66%)	505 (70%)	501 (59%)
Black or African American	433 (22%)	140 (19%)	161 (19%)
Asian	27 (1.3%)	12 (1.7%)	11 (1.3%)
American Indian or Alaska Native	3 (0.1%)	6 (0.8%)	0 (0%)
Native Hawaiian or Other Pacific Islander	4 (0.2%)	3 (0.4%)	0 (0%)
Other	200 (10%)	58 (8%)	179 (20.6%)
Vitals and Laboratory Values			
Pulse	81 (71, 93)	82 (70, 94)	81 (71, 93)
SBP	123 (111, 139)	125 (113, 140)	131 (116, 146)
DBP	68 (59, 78)	69 (60, 79)	69 (61, 79)
SpO2	98 (96, 100)	98 (96, 100)	98 (97, 99)
Hemoglobin	11.50 (9.20, 13.50)	10.80 (8.67, 13.00)	11.05 (8.30, 13.30)
Platelets	228 (165, 297)	219 (157, 290)	226 (172, 285)
INR	1.04 (0.97, 1.20)	1.07 (0.99, 1.26)	1.12 (1.05, 1.29)
ALT	18 (12, 29)	18 (13, 29)	31 (24, 43)
AST	23 (18, 36)	25 (19, 41)	26 (20, 40)
Total Bilirubin	0.40 (0.28, 0.70)	0.40 (0.30, 0.80)	0.60 (0.40, 0.90)

[1] Median (IQR); n (%)

The training and test sets were from the same center and temporally separated, with similar demographic and laboratory values at presentation. The external validation was from a separate center and had increased serological tests for liver dysfunction (ALT, AST, total bilirubin), less proportion of the population identifying as Caucasian (59% versus 66-70%), and slightly more abnormal INR (1.12 versus 1.04-1.07). The components of the GBS and Oakland Score for patients in the external validation dataset is shown. (Table 10)

Table 10: Components of the Glasgow Blatchford Score and Oakland Score in the external validation dataset.

Glasgow Blatchford Score	External Validation (N=849)
	Mean (SD)
Total Score	6.2 (0.15)
Blood Urea Nitrogen	1.9 (0.06)
Hemoglobin	2.9 (0.09)
Systolic Blood Pressure	0.25 (0.02)
Pulse	0.14 (0.01)
	Proportion (%)
Melena	261 (31%)
Syncope	57 (6.7%)
Hepatic Disease	120 (14%)
Cardiac Failure	132 (15%/)

Oakland Score	External Validation (N=849)
	Mean (SD)
Total Score	16.8 (0.24)
Pulse	1.2 (0.03)
Systolic Blood Pressure	2.5 (0.04)
Hemoglobin	11.1 (0.22)
Age	1.4 (0.02)
	Proportion (%)
Gender	470 (65%)
Previous LGIB	82 (11%)

External Validation

The AUC for the deep learning model (AUC=0.92, 0.90-0.94) was similar to the gradient boosted machine learning model (AUC=0.92, 0.90-0.94; p=0.93) and was higher than GBS (0.88, 0.85-0.90; p=0.005) and Oakland (0.89, 0.87-0.91; p=0.057) (Table 11).

Table 11: Area under the receiver operating curve (AUC) for the gradient boosted decision tree machine learning model vs. Glasgow-Blatchford Score (GBS) and Oakland Score.

	Deep Learning Model AUC (95% CI)	Machine Learning Model AUC (95% CI)	P-value	GBS AUC (95% CI)	P-value	Oakland AUC (95% CI)	P-value
External Validation (N = 849)	0.92 (0.90-0.94)	0.92 (0.90-0.94)	0.93	0.88 (0.85-0.90)	0.005	0.89 (0.87-0.91)	0.057

At 100% sensitivity, the machine learning model has higher specificity than Oakland=4 (21% vs 2%, p<0.001); no GBS achieved 100% sensitivity in external validation dataset (Table 12). At 99% sensitivity, the machine learning model has higher specificity (41.6%) than GBS=0 (20.5%, p<0.001) and Oakland=8 (19.7%, p<0.001) (Table 4). At the 99% sensitivity cutoff, the machine learning model correctly identified more very-low-risk patients (N=214) among the 849 patients in this cohort than GBS=0 (N=106) and Oakland=8 (N=104).

Table 12: Specificities of the gradient boosted decision tree machine learning model vs. Glasgow Blatchford Score and Oakland Score at high sensitivity cutoffs to identify very-low-risk patients in the external validation dataset (N = 849).

	Sensitivity	Specificity	P-Value
Deep Learning Model	100%	17.8%	
Machine Learning Model	100%	11.2%	<0.001
Oakland = 4	100%	2%	
Deep Learning Model	99%	41.6%	
Machine Learning Model	99%	30.6%	<0.001
GBS = 0	99%	20.5%	
Oakland = 8	99%	19.7%	

Calibration of the deep learning model (0.11) was equivalent to the gradient boosted decision tree model (0.11) and better than the GBS (0.14) and Oakland models (0.13) by Brier scores (Figure 10).

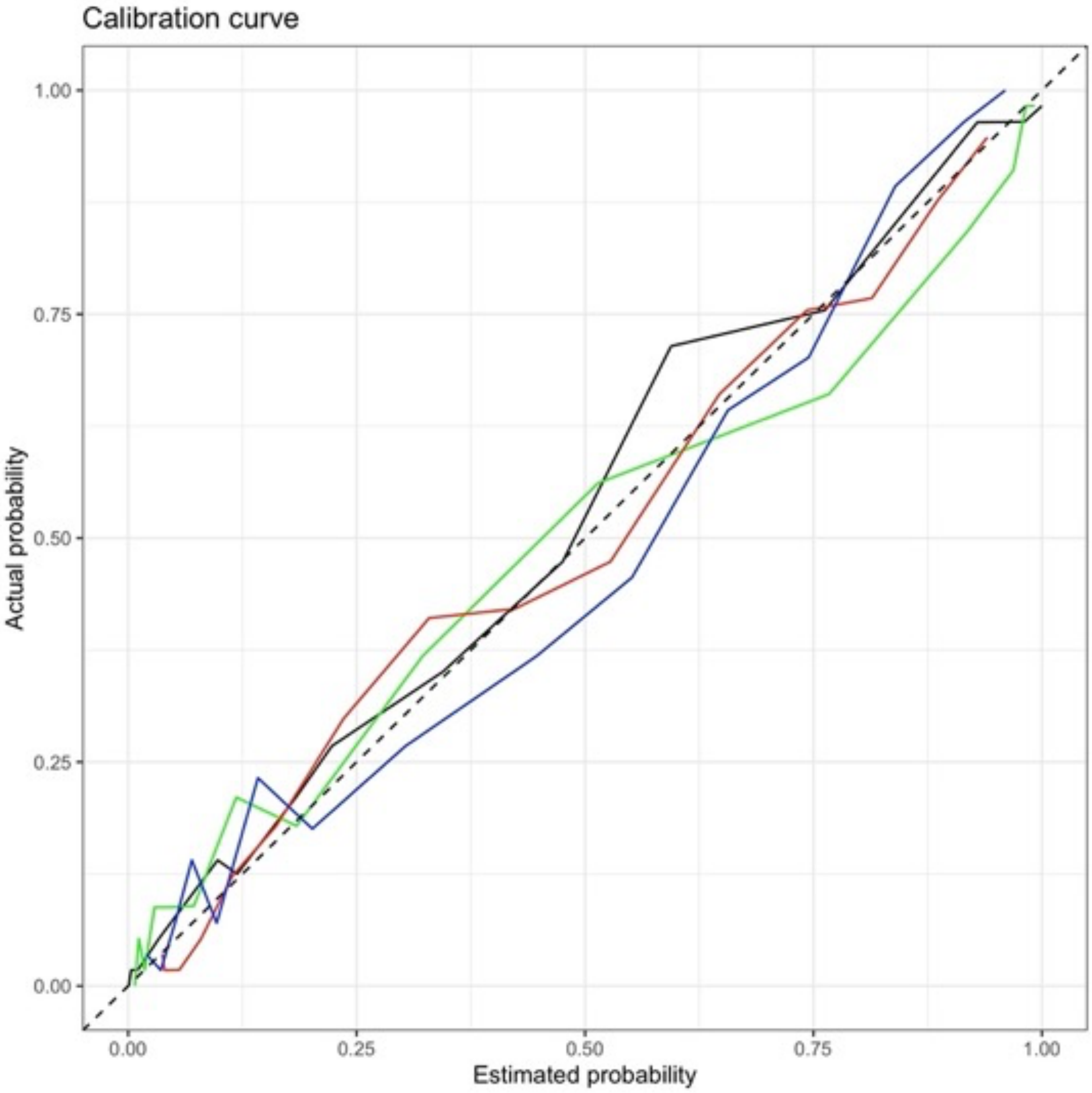


Figure 10: Calibration curves for the neural network model (black, Brier score 0.11), gradient boosted model (green, Brier score 0.11), Glasgow-Blatchford Score (red, Brier score 0.14), and Oakland Score (blue, Brier score 0.13) for external validation.

Discussion

This is the first study to develop and validate an EHR-based deep learning model that provides excellent performance in predicting very-low-risk patients presenting with acute GIB who are eligible for discharge from the emergency room without admission to hospital. This deep learning model demonstrates similar overall performance to the gradient boosted decision tree model and superior overall performance as compared to the two guideline-recommended clinical risk scores, the GBS and the Oakland Score. Importantly, at cutoffs (sensitivity of 99% and 100%) designed to avoid falsely labeling patients who will die or require transfusion or intervention as very-low-risk, the neural network model identifies more patients who are eligible for discharge and outpatient management than the gradient boosted decision tree model and currently recommended clinical risk scores. The model was developed specifically on data collected and available in the EHR within the first four hours of presentation, which allows the potential integration of the model in real time to provide decision support. Additionally, the score can be automatically calculated and does not require manual data extraction and input from the provider due to the use of only structured datafields from the EHR.

Strengths

While many machine learning models have been developed, the gap between development and implementation is hampered by the choice of dataflow, absence of external validation and calibration, and no consideration for where or how the tool will integrate into and enhance provider workflow. For example, 2 recent studies of EHR-based machine learning models for acute GIB have examined the use of machine learning tools to predict in-intensive-care-unit mortality[93] and need for red-blood-cell transfusion [94] for patients in the intensive care unit. These models may have more limited utility given the fact that most patients with GIB aren't admitted to the intensive care unit and many important care decisions (e.g., admission vs. discharge, level of care if admitted, initial resuscitation and transfusion requirements, initial medications for GIB, timing of endoscopy) are made prior to placement in the intensive care unit. In contrast, our study specifically targets the guideline-recommended risk stratification triage in the emergency department. We provide a comprehensive scheme for developing and validating an EHR-based deep learning and machine learning tools that are designed for integration into the assessment of patients with acute GIB in the emergency room in 4 hours from presentation and can be used to help decide whether very-low-risk patients can be discharged for outpatient management. Since these models are developed using available structured datafields available within 4 hours of the patients' presentation, it can automatically extract the necessary data from the patient record and automatically calculate a risk score after the data are processed using the gradient boosted tree model. This study demonstrates robust performance on a temporally separate patient cohort presenting to the same centers and also on external validation in a separate hospital using the same EHR system. In the era of learning health systems, this study suggests that machine learning with EHR data has the potential of scalability not only at one center but multiple centers. The development of our model in the most widely used EHR system in the U.S. may be useful in generalizing this algorithmic approach to other healthcare systems.

Limitations

Complete cases analysis was performed, limiting the use of all training examples. Missingness in EHR data, especially laboratory data, is of particular concern since practice patterns may be biased by experience, seniority, and provider perception. In order to allay those concerns in this analysis, only patients with measured values for laboratory values and vital signs were included in the analysis. In future studies, new ways of modeling missing data and comparing the performance of various approaches of data imputation will be important to increase representation of all patient data. The findings of this study, while promising, requires prospective implementation in live ED provider workflow to evaluate real world performance. Importantly, feasibility and usability studies that assess provider acceptance is critical. Interpretability of the machine learning tool is also an aspect that needs to be studied, particularly due to the role of the tool in assisting clinicians' decision making.

Development and Validation of Electronic Health Record Phenotypes for Acute Gastrointestinal Bleeding

Introduction

Acute gastrointestinal bleeding (GIB) is the most common gastrointestinal diagnosis requiring hospital admission in the United States.[69] Guidelines for upper and lower GIB recommend risk stratification of patients, including the use of risk assessment scores.[7,9,95] Although many risk stratification tools have been developed and validated, they are not commonly used in real-world clinical practice partly because providers must manually enter a variety of variables into the scoring system. Widespread electronic health record (EHR) adoption makes it possible to automatically deploy risk stratification scores within the clinical workflow for acute GIB; however, in order to embed risk stratification models into EHRs and deploy them in real time, patients must first be correctly identified. This process of accurately identifying patients, called phenotyping, is used for any study that seeks to reliably group patients with a specific diagnosis or condition, from surveillance studies to comparative effectiveness research. Phenotyping is typically the first step in developing and validating risk stratification models within EHRs.[96] Such processes have been used to improve the accuracy of case definition of inflammatory bowel disease patients as Crohn's disease or ulcerative colitis, as well as to facilitate clinical trial recruitment and deploy randomized controlled trials.[97,98]

Unlike many conditions that require multiple elements of the record (laboratory testing, reported symptoms, and biometrics such as vital signs) for diagnosis, acute GIB is a condition that can be directly and clearly identified using a limited number of terms by patient report or provider evaluation. To our knowledge, no previous study has explored the early identification of patients with acute GIB with an EHR-based model or its implementation within standard EHR workflow.

Phenotypes utilize both structured and unstructured data and are typically used retrospectively after the clinical encounters have ended (e.g., ICD codes are a popular component of phenotypes). If the goal is to identify patients that would benefit from predictive models tailored for a particular condition, EHR phenotypes must use data elements generated during the visit. The Systematized

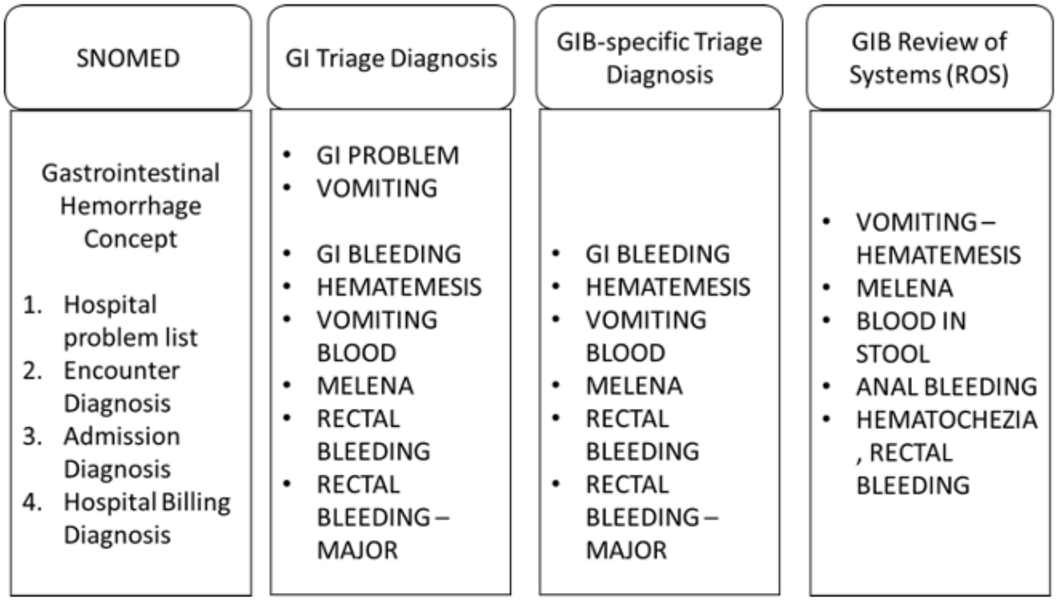
Nomenclature of Medicine (SNOMED) is an international comprehensive clinical terminology that is the standard for encoding patient conditions in the EHR. Other approaches for phenotyping can be decision rules using specific data elements (e.g., triage diagnosis) or a machine learning (ML) approach to utilize unstructured clinical text through natural language processing (NLP). ML models use computational modeling to learn from data and their performance can improve with an increasing amount of data. NLP is a set of tools used to extract data from narrative text and uses syntactic processing, the order and arrangement of words in sentences and phrases, and semantic analysis to capture the meaning of the text. The empirical performance of these tools could provide useful comparisons to decision rules. For prediction, however, a "screening" phenotype using data elements likely to be entered close to real time may be better than a tool that may be delayed (e.g. due to delays in note writing).[25]

Our study aimed to accurately identify patients with acute GIB reported or witnessed in the emergency department, such that the identification of this phenotype can occur in real time in the EHR to subsequently launch predictive models for risk stratification. We chose to use the current standard for phenotyping patients, SNOMED, as the comparator, even though SNOMED does not provide real-time phenotyping in the emergency department.

Methods

We began with creating a sensitive data mart, a patient dataset selected using specified criteria, to screen for all patients presenting with GIB from 2014 to 2017 in the Yale New Haven Health System electronic health record (Epic, Verona, WI, USA). Creation of a sensitive data mart allows exclusion of patients with no evidence of the phenotype and to adequately handle the volume, heterogeneity, and velocity of data.[99] To create the data mart, we screened for patients with data that suggested the phenotype of overt gastrointestinal bleeding. In order to maximize the capture of relevant data, we defined the process by which patients were evaluated in the emergency department and common time periods for data entry. We also identified points at which a diagnosis would be entered in the electronic health record throughout the hospital stay: hospital problem list, encounter diagnosis, admission diagnosis, and hospital billing diagnosis.

There were 4 categories of screening criteria, which were selected based on existing identifiers in the EHR used to denote gastrointestinal bleeding (Figure 11).



SNOMED	GI Triage Diagnosis	GIB-specific Triage Diagnosis	GIB Review of Systems (ROS)
Gastrointestinal Hemorrhage Concept 1. Hospital problem list 2. Encounter Diagnosis 3. Admission Diagnosis 4. Hospital Billing Diagnosis	• GI PROBLEM • VOMITING • GI BLEEDING • HEMATEMESIS • VOMITING BLOOD • MELENA • RECTAL BLEEDING • RECTAL BLEEDING – MAJOR	• GI BLEEDING • HEMATEMESIS • VOMITING BLOOD • MELENA • RECTAL BLEEDING • RECTAL BLEEDING – MAJOR	• VOMITING – HEMATEMESIS • MELENA • BLOOD IN STOOL • ANAL BLEEDING • HEMATOCHEZIA, RECTAL BLEEDING

Figure 11: Screening criteria used to create data mart for acute gastrointestinal bleeding.

Development of EHR GIB Phenotypes

We developed rule-based algorithms and two machine-learning based algorithms, one using syntactic NLP analysis and the other using a Bidirectional Encoder Representation from Transformers (BERT) neural network NLP model and compared their performance to the SNOMED-only classification (Figure 12).

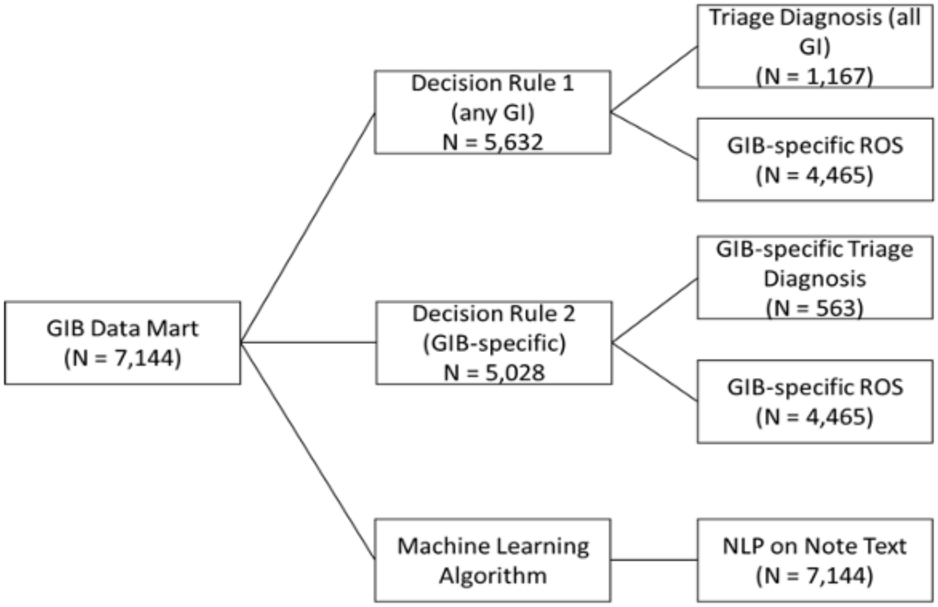


Figure 12: Rule-based algorithms and machine learning (Natural Language Processing) algorithm.

To determine the specific data elements included in the rule-based algorithm, we analyzed the clinical workflow to identify two points where relevant diagnosis or symptom data was entered into the EHR. The first point of data entry was at triage, where a nurse selected presenting diagnoses from a drop-down list of pre-specified diagnoses. We reviewed all triage diagnoses to identify any that were gastrointestinal diagnoses. The second data entry point was the review of systems section in the note template used for all ED patients, which contains elements referring to overt gastrointestinal bleeding. Both the triage diagnosis and review of systems provide structured data fields that can be used to identify patients as either 0 (not present) or 1 (present). We hypothesized that terms specific to gastrointestinal bleeding may have improved performance and therefore created two decision rules. Decision rule 1 was positive if the ROS field (e.g., hematemesis, blood in stool) was positive or if any GI triage term was positive. Decision rule 2 was positive if the ROS field was positive or GIB triage terms were positive.

The syntax-based NLP approach includes the preprocessing of the unstructured text in notes written by physician providers in the Emergency Department using ScispaCy, a Python software library used for advanced NLP that allows for breaking down text into smaller unique parts, which is a strategy applied to other biomedical text data.[100, 101,102] Classification was performed using random forest, support vector machine, and elastic net classifiers.

The NLP approach to capture meaning, or sematic information, from physician notes was the Bidirectional Encoder Representation from Transformers (BERT) model.[103] The BERT model is a neural network Transformer based model that generates embeddings for sentences to capture meaning. We used Clinical BERT, a variant of BERT fine-tuned on biomedical and clinical text corpora (MIMIC-III and PubMed).[104]

We performed a sensitivity analysis for hematemesis and/or melena and hematochezia. For hematemesis or melena decision rule 2 was modified to only include triage diagnoses GI Bleeding, hematemesis, melena, vomiting blood and review of systems elements of melena and hematemesis. For hematochezia decision rule 2 was modified to only include triage diagnoses GI Bleeding, major rectal bleeding, rectal bleeding, and review of systems elements of blood in stool, anal bleeding, rectal bleeding, and hematochezia.

Evaluation of Phenotype

To assess the performance of the different phenotypes, we performed manual note review to create a gold standard. Two clinical domain experts (DS and CT) reviewed the medical records of all patients and classified each patient as having the phenotype or not based on expert opinion and a prespecified structured evaluation (Table 13). We further categorized the acute bleeding by symptoms, either hematemesis/melena or hematochezia.

Table 13: Gold-Standard Strategy to Label Encounters

Steps to label the Gold Standard: Acute Gastrointestinal Bleeding in the ED

1. Review of the ED Provider Note
2. Any text that identifies acute gastrointestinal bleeding for
    a. Hematemesis: e.g. “hematemesis”, “vomiting blood”
    b. Melena: e.g. “dark stool”, “black stool”, “tarry stool”, “melena”
    c. Hematochezia: e.g. “blood in toilet bowl”, “blood in stool”, “bright red blood in the stool”
3. Patient report or physical exam findings were considered equally valid
4. Exclude patients with other reasons for hematemesis – e.g. epistaxis

Training and Validation Datasets

The total number of encounters was temporally divided into training (70%, from 9/2014 to 7/2016) and validation (30%, 7/2016 to 5/2017) sets. Internal validation was performed with ten replications of tenfold cross-validation across the training set for decision rules and the syntax-based natural language processing. For each cross-validation split, the McNemar’s test was performed for sensitivity, specificity, and positive predictive value comparing SNOMED to the decision rule 1, decision rule 2, and syntax-based NLP with random forest, support vector machine, and elastic net classifiers. Unfortunately, due to computational constraints we did not perform the internal validation on the BERT neural network model. External validation was performed directly using the held-out validation set. The primary metric for performance was the positive predictive value, and secondary metrics for performance the sensitivity and specificity. A high PPV indicates that a high proportion of the patients identified with acute GIB are true cases, but ideally would also have a high sensitivity to identify a high proportion of all true cases of acute GIB. There is no clear performance threshold for PPV, but PPV>75% has been considered acceptable and reported for EHR phenotypes.[105-110]

Sensitivity Analysis by Bleeding Etiology

Pre-defined sensitivity analysis was performed in the same external validation cohort to predict either hematemesis and/or melena or hematochezia. These are two clinically distinct symptom complexes that may indicate an upper gastrointestinal tract or lower gastrointestinal tract source, respectively.

Statistical Methodology

McNemar's test was used to compare sensitivity, specificity, and PPV for each iteration of internal validation for SNOMED, decision rules, and NLP tools. For the ten replications of tenfold internal validation, the median and range of p-values for the McNemar's test are presented (Tables 14 and 15).

Table 14: Internal Validation with Tenfold Cross-Validation for 10 iterations: SNOMED versus Decision Rules.

Training (70%) N = 7144 Internal Validation: Tenfold Cross-Validation for 10 iterations					
Phenotype and Performance Characteristics					
	Performance Algorithm (with 99% confidence interval)				
	SNOMED Codes (reference) Problem List Encounter Diagnosis Billing Diagnosis	Decision Rule 1: All GI Triage Terms + ROS fields	P-Value (Median, Range)	Decision Rule 2: GI Bleed-Specific Triage terms + ROS Fields	P-Value (Median, Range)
PPV	74% (0.740 – 0.746)	85% (0.849 – 0.855)	<0.0001 (0 – 0.002)	91% (0.907 – 0.913)	<0.0001
Sensitivity	61% (0.606 – 0.618)	91% (0.904 – 0.910)	<0.0001	88% (0.882 – 0.889)	<0.0001
Specificity	39% (0.382 – 0.399)	55% (0.536 – 0.555)	0.036 (0 – 0.50)	75% (0.740 – 0.757)	<0.0001

Table 15: Internal Validation with Tenfold Cross-Validation for 10 iterations: SNOMED versus NLP approaches.

Training (70%) N = 7144
Internal Validation: Tenfold Cross-Validation for 10 iterations
Phenotype and Performance Characteristics

	Performance Algorithm (with 99% confidence interval)						
	SNOMED Codes (reference) Problem List Encounter Diagnosis Billing Diagnosis	Syntax-based NLP with Random Forest	P-Value (Median, Range)	Syntax-based NLP with Support Vector Machines	P-Value (Median, Range)	Syntax-based NLP with Elastic Net	P-Value (Median, Range)
PPV	74% (0.740 – 0.746)	86% (0.857 – 0.862)	<0.0001	88% (0.880 – 0.885)	<0.0001	89% (0.888 – 0.893)	<0.0001
Sensitivity	61% (0.606 – 0.618)	97% (0.969 – 0.973)	<0.0001	94% (0.937 – 0.942)	<0.0001	90% (0.895 – 0.902)	<0.0001
Specificity	39% (0.382 – 0.399)	54% (0.533 – 0.551)	0.005 (0-0.16)	64% (0.630 – 0.649)	<0.0001 (0-0.001)	68% (0.672 – 0.692)	<0.0001 (0-0.0003)

We considered ten replications of tenfold cross-validation to account for variabilities of random splitting, so that we could generate a more robust understanding of potential performance for each of the approaches. McNemar's test was also applied for the external validation dataset (Table 16).

Table 16: External Validation of NLP phenotypes on temporally separate cohort of patients.

External Validation (30%) N = 2988

Phenotype and Performance Characteristics

	Performance Algorithm (with 99% confidence interval)			
	SNOMED Codes (reference) Problem List Encounter Diagnosis Billing Diagnosis	Syntax-based NLP with Random Forest	Syntax-based NLP with Support Vector Machines	Syntax-based NLP with Elastic Net
PPV	74% (0.71 – 0.76)	80%* (0.78 – 0.82)	81%* (0.79 – 0.83)	83%* (0.81 – 0.86)
Sensitivity	61% (0.59 – 0.64)	98% * (0.96 – 0.98)	95%* (0.93 – 0.96)	85%* (0.83 – 0.87)
Specificity	38% (0.34 – 0.43)	51%* (0.46 – 0.54)	54%* (0.50 – 0.58)	65%* (0.61 – 0.69)

P<0.001 compared to SNOMED (baseline)

PPV was considered the primary metric for performance with goal PPV >75%. We also predefined comparisons between the baseline (SNOMED) and each of the phenotyping approaches (decision rule 1, decision rule 2, and the two NLP-based approaches). We corrected using the Bonferroni correction and defined significance as *P*<0.01; we present the corresponding 99% confidence intervals.

Results:

Performance of Decision Rules

Internal Validation with Tenfold Cross-Validation for Ten Iterations

Decision rules 1, 2, and the syntax-based NLP tools had better PPVs than SNOMED codes in identifying patients with acute bleeding. The syntax-based NLP tool with random forest classifier had the highest sensitivity (0.97, 99% CI: 0.969-0.973) and was better than SNOMED (0.61, 99% CI: 0.606-0.618) with median P value <0.0001. Decision Rule 1, Decision Rule 2 (0.88, 99% CI: 0.882-0.889), and the syntax-based NLP tool with elastic net and support vector machine classifiers also had higher sensitivity than SNOMED (median P value <0.0001). Decision Rule 2 had the highest specificity (0.75, 99% CI: 0.740-0.757) and was better than SNOMED (0.39, 99% CI: 0.382-0.399) with median P value <0.0001). (Table 14, Table 15)

External Validation

The PPV of decision rules (0.78, 99% CI: 0.75-0.80 for decision rule 1, 0.85, 99% CI: 0.83-0.87 for decision rule 2) were increased compared to SNOMED (0.69, 99% CI: 0.66-0.72; *P*<0.001). Syntax-based NLP with the elastic net classifier (PPV 0.83, 99% CI: 0.81-0.86) and BERT neural network model (PPV 0.84, 99% CI: 0.82-0.86) also were increased compared to SNOMED (P<0.001). The sensitivity of decision rules (0.90, 99% CI: 0.88-0.92 for decision rule 1, 0.87, 99% CI: 0.85-0.89 for decision rule 2), syntax-based NLP with the elastic net classifier (0.85, 99% CI: 0.83-0.87), and the BERT neural network NLP model (0.93, 99% CI: 0.92-0.95) are increased compared to SNOMED codes (0.59, 99% CI: 0.57-0.62;*P*<0.001). For specificity, SNOMED codes (0.45, 99% CI: 0.42-0.48) is worse than decision rule 2 (0.69, 99% CI: 0.65-0.73;*P*<0.001), syntax-based NLP with the elastic net classifier (0.65, 99% CI: 0.61-0.69), and BERT neural network NLP model (0.63, 99% CI: 0.59-0.67) but similar to decision rule 1 (0.47, 99% CI: 0.43-0.51;*P*=0.87). (Table 16)

Sensitivity Analysis

Sensitivity Analysis based on Type of Bleeding

On external validation for hematemesis and/or melena decision rule and NLP algorithms had higher PPV than SNOMED codes ($p<0.001$). The BERT Neural Network NLP model had a similar sensitivity to SNOMED codes ($p=0.77$), while the syntax-based NLP models and the modified decision rule had lower sensitivity ($p<0.001$). For hematochezia alone, the decision rule and NLP algorithms had higher PPV than SNOMED codes ($p<0.001$). (Table 16)

Table 16: Sensitivity Analysis of Hematemesis and/or Melena and Hematochezia in the External Validation Group.

Hematemesis and/or Melena Validation (30%) N = 650/2988				
Phenotype and Performance Characteristics				
	Performance Algorithm (with 99% confidence interval)			
	SNOMED Codes (reference)	Decision Rule 2 Modified: Upper GI Bleed-Specific Triage Terms + ROS fields	Syntax-based NLP with Random Forest	BERT Neural Network NLP
PPV	29% (0.26 – 0.31)	82%* (0.76 – 0.87)	85%* (0.81–0.90)	78%* (0.74-0.83)
Sensitivity	71% (0.67 – 0.75)	34%* (0.30 – 0.39)	46%* (0.42–0.51)	72% (0.68-0.76)
Specificity	46% (0.44 – 0.49)	98%* (0.97 – 0.98)	98%* (0.97–0.98)	94%* (0.93-0.95)
Hematochezia Validation (30%) N = 1316/2988				
Phenotype and Performance Characteristics				
	Performance Algorithm (with 99% confidence interval)			
	SNOMED Codes (reference)	Decision Rule 2 Modified: Lower GI Bleed-Specific Triage Terms + ROS fields	Syntax-based NLP with Support Vector Machines	BERT Neural Networks NLP
PPV	42% (0.39 – 0.45)	72%* (0.69 – 0.74)	86%* (0.83 – 0.88)	84%* (0.82 – 0.87)
Sensitivity	55% (0.51 – 0.58)	87%* (0.85 – 0.89)	82%* (0.79 – 0.84)	90%* (0.88 – 0.92)
Specificity	40% (0.37 – 0.43)	73%* (0.70 – 0.75)	89%* (0.87 – 0.91)	87%* (0.84 – 0.89)

P<0.001 compared to SNOMED (baseline)

Discussion

This is the first study to develop an EHR phenotype for identifying acute gastrointestinal bleeding in real time. Prompt identification of individuals with acute gastrointestinal bleed in the emergency room is an important first step in order to deploy risk scores that would inform determine level of care and clinical management decisions. (Figure 13)

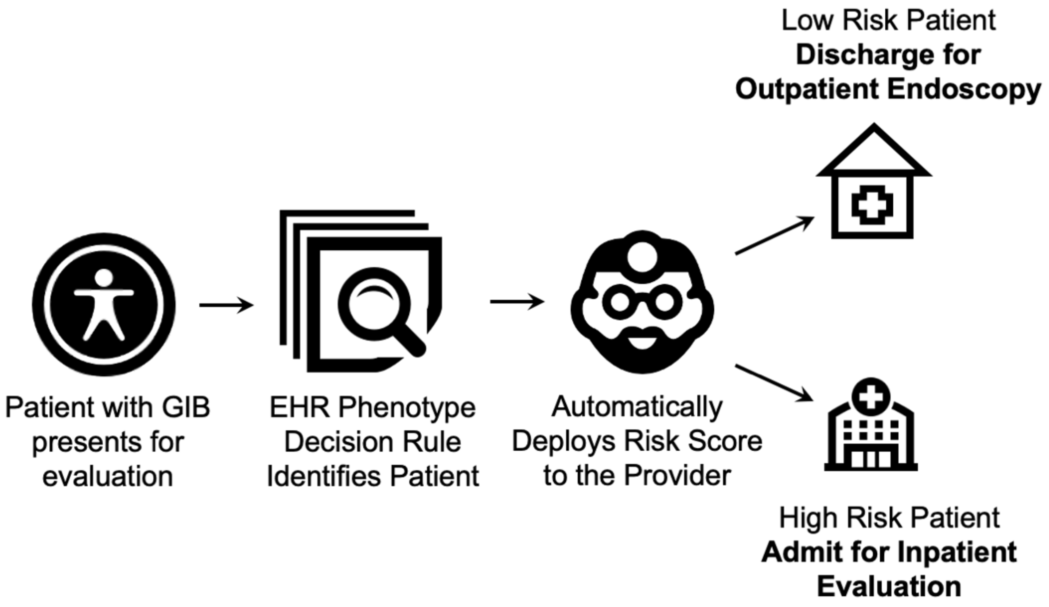

Low Risk Patient
Discharge for Outpatient Endoscopy
Patient with GIB presents for evaluation
EHR Phenotype Decision Rule Identifies Patient
Automatically Deploys Risk Score to the Provider
High Risk Patient
Admit for Inpatient Evaluation

Figure 13: Schematic showing how the electronic health record (EHR) phenotyping rule fits into the provider workflow to find patients with gastrointestinal bleeding and deploy risk scores to assist decision making.

We found that the automated decision rule that combined bleed-specific terms at initial triage by a nurse and bleed-specific terms in the emergency department provider's review of systems had the highest PPV (85%) on external validation with a sensitivity of 84% to identify patients presenting with acute GIB in the emergency department. In practice, this would generate an alert that would deploy a risk stratification tool for acute GIB through the electronic health record to the provider. A proposed workflow would be the following: a patient presents to the emergency department with acute gastrointestinal bleeding and is identified with the decision rule from a triage GIB-specific diagnosis of "Vomiting Blood." The patient would then be flagged and his vital signs, laboratory values, and medical history elements from the electronic health record would be extracted and run through a risk stratification algorithm that would identify very low risk patients with a pre-specified threshold (e.g. machine learning-based risk stratification tools).[13,14] This algorithm would then prompt an alert that would be delivered to the provider to identify very low risk patients who could be discharged.

We chose the decision rule because it identifies patients who actually have acute GIB, which means we want a low false positive rate (high specificity and positive predictive value). Sensitivity should also be as high as possible to ensure that most patients with acute GIB are detected. A positive predictive value of >75% has been reported as "high," and the PPV for the highest performing decision rule was 91% on internal validation and 85% on external validation.[105-110]

Traditionally no method exists to identify patients with gastrointestinal bleeding other than using diagnostic codes (ICD-9, ICD-10) from the billing diagnosis list. By definition, these diagnostic codes are useless for detecting patients presenting with acute GIB at the point of care because they generally are not entered into the EHR until much later in the hospital course, well after initial identification and risk assessment is needed. Additionally, ICD codes may not detect all patients who have bleeding, as it may be reported but not coded in their billing diagnosis. SNOMED may be better than diagnostic codes alone—it is a terminology that includes both billing codes and other related text, and is updated every quarter.[111,112] Our decision rule outperforms the baseline SNOMED methodology even when sampled over multiple areas in the EHR, including the billing diagnosis list. Interestingly, using multiple strategies of NLP over the entire text of the emergency department provider notes does not result in markedly improved performance over a decision rule for identifying patients with acute GIB, though it appears to have a slight benefit in the sensitivity analysis for hematochezia. Unfortunately, the NLP tools are not applicable to real-time phenotyping, since the availability of the entire text is delayed since providers may choose to complete documentation up to 24 hours after the actual visit.

The NLP tools had similar performance to decision rules for detecting overall acute bleeding. On sensitivity analysis, NLP tools performed similarly to the modified decision rule for melena or hematemesis, and better than the modified decision rule for hematochezia. One advantage of NLP-based tools over decision rules is flexibility in the ability to handle additional covariates, including demographics and clinical information. In our evaluation, we incorporated age and gender into the features for the syntax-based NLP algorithm, but the results were similar. Future work is needed to systematically evaluate the additional benefits of incorporating other types of variables into NLP classification.

Strengths

This study compares the performance of existing automated methods (SNOMED) across multiple time points in the EHR in identifying acute GIB at admission. This excludes GIB after admission while already hospitalized for another condition, and provides the best data possible given the constraints of using only structured datafields. Prompt identification of individuals with acute GIB would allow for the automatic provision of risk stratification scores for providers to guide appropriate triaging and clinical decision-making.

Limitations

We did not review patients who did not have the specific SNOMED code for GIB, did not have a GI-related triage problem, and did not have a positive review of systems for GIB. We believe it is reasonable to assume that these patients likely present with another primary issue and without any clinically significant GIB. Risk stratification scores for acute gastrointestinal bleeding are typically used for patients who present with GIB as the chief and acute complaint, and clinical decisions, such as admission or hospital-based interventions, need to be made early after presentation.

This phenotype was developed for a specific center with a local workflow, including the availability of a triage nurse with structured datafields for triage diagnosis. Patients outside this cohort (with negative SNOMED, none of the triage terms, and no ROS positivity) were not reviewed, which limits its applicability to all-comers in the emergency department. However, we believe patients without any GI symptoms or signs at triage or during emergency department evaluation and without any evidence of GIB on diagnostic codes or the other elements of SNOMED are very unlikely to have presented with clinically significant acute GIB. Changes in coding (e.g., from ICD-9 to ICD-10) and temporal shifts in treatment options, patient epidemiology, hospital utilization, and risk shifts can all decrease the performance of these phenotypes in identifying patients of interest.

Preliminary Results from Real-Time Implementation to Prompt an Acute Gastrointestinal Bleeding-specific Care Pathway

From the results that suggested that the decision rule is a robust method of identifying patients with acute GIB, we implemented the phenotype as a way to identify patients who may benefit from the utilization of a clinical care pathway for acute gastrointestinal bleeding.

Introduction

Learning health systems can be defined as a broader system in which science, informatics, incentives, and culture align to integrate best practices into the health delivery process for improvement and innovation.[113] The information and knowledge are thought to be captured as a by-product of clinical care delivery, which are ideally longitudinally tracked with readily available outcomes.[114] In order to create clinical decision tool available in real time that utilizes the data generated during clinical care and integrates best evidence-based practice, one approach is to utilize clinical pathways integrated into the electronic health record.

Clinical pathways, also known as critical or integrated care pathways, comprise a specific approach that seeks to align evidence-based practices into the care workflow in a standardized manner.[115] These pathways were first used in the 1980’s in response to changes in reimbursement policy turned the focus away from high volumes to improving patient outcomes.[116] With the advent of the electronic health record, clinical pathways have been adapted from a paper-based system to become integrated within the electronic health record workflow with additional functionalities.[117] These pathways now integrate order sets based on specific recommendations, clinical risk score calculators, and visualization of available therapeutic options.

Yale-New Haven Health System has recently made a significant investment in the Care Signatures Initiative, which has designed pathways for a variety of clinical conditions using stakeholders across the different departments and disciplines. These pathways have been integrated into the electronic health record via a visualization available for providers to reference during the course of patient care. While adoption across different contexts may require modifications given resource availability, population shift, and workflow differences, clinical pathways are usually designed around best practices for specific conditions that apply across all health settings. However, these digital pathways by themselves have low provider uptake and are insufficient for effective clinical decision support. They require a robust delivery strategy that automates the steps of identifying the right patients using elements in the electronic health record so that the pathways can be deployed to the right person, for the right patient, at the right time.[118]

The Care Signatures Pathway for Acute Gastrointestinal Bleeding provides a workflow to assess, triage, manage, and formulate a follow-up plan for patients presenting with signs and symptoms of acute gastrointestinal bleeding. As part of the pathway, providers are given best practice guidelines for aspects of history and physical examination helpful for assessing the patient risk for needing hospital-based intervention, links to order sets to facilitate care (such as a consult link, intravenous fluids, and blood transfusion order sets), and considerations for patients with special conditions such as anticoagulant or antithrombotic medications or patients at risk of portal hypertensive bleeding.

Methods

Given the potential utility of the Care Signatures Pathway for Acute Gastrointestinal Bleeding, the decision rule phenotype was implemented in the Epic system such that it would flag the use of specific pathways for patients who fulfilled the decision rule criteria. Implementation of the pathways occurred on March 5th, 2021 for inpatients, on June 22nd, 2021 for the emergency department, and the decision rule to suggest the pathway was implemented in November 19th, 2021. The suggestion was available for all providers accessing the patient chart at all times during the hospital stay.

Results

The rate of inpatient usage before the suggestion system was implemented was approximately 9.4 unique encounters per week, and emergency department usage was 2 per week. After implementation, inpatient usage was up to 11.2 unique encounters per week while emergency department usage was 2.7 unique encounters per week. Notably, the suggestion system accounted for 21% of inpatient encounters for which the clinical pathway was utilized and 34% of emergency department encounters. (Figure 14)

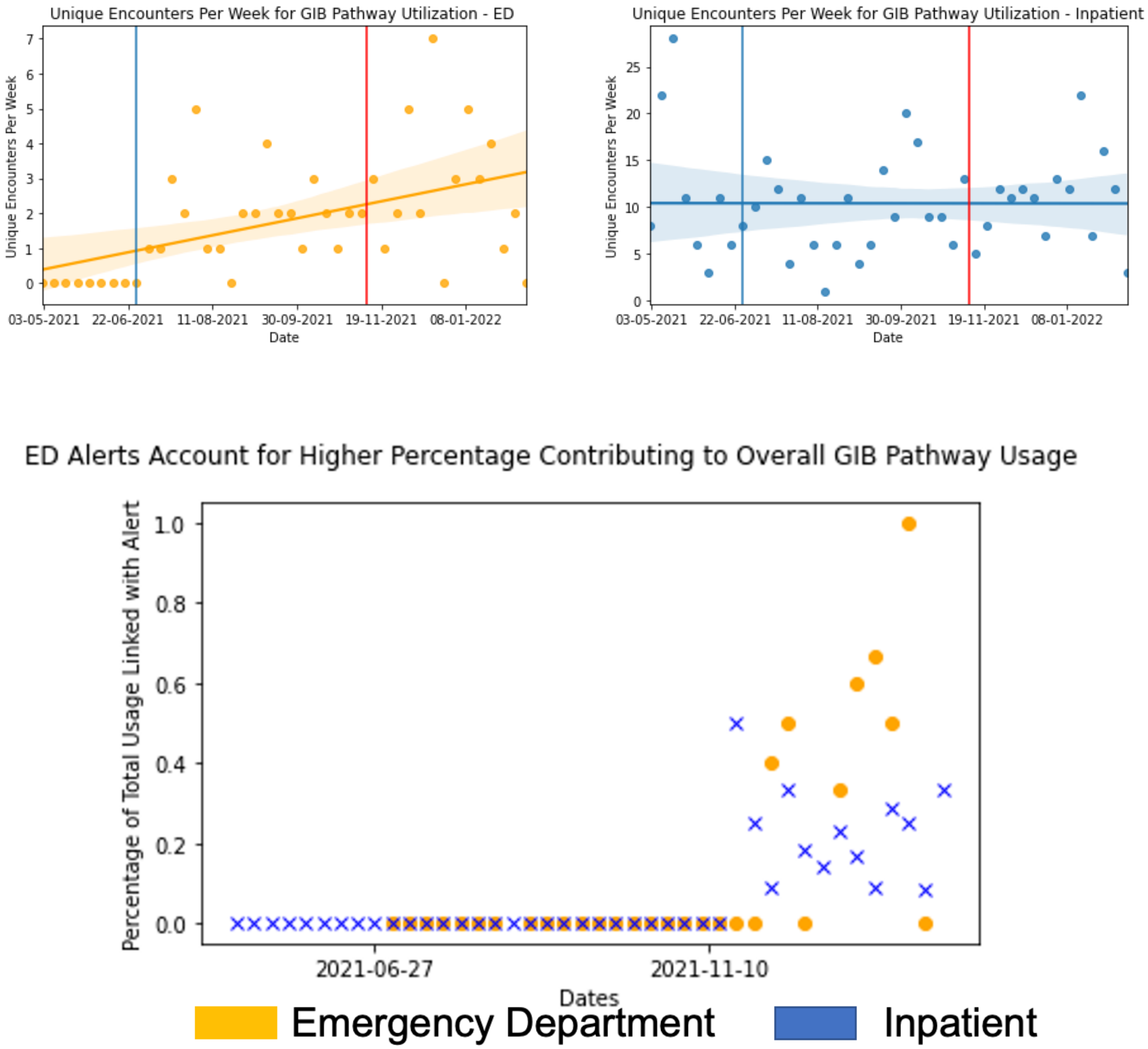


Figure 14: Trends of Clinical Care Pathway Utilization for Acute Gastrointestinal Bleeding after Introduction of the ED GIB Pathway and then After Introduction of the EHR Phenotype.

Preliminary Conclusions

Implementation of a suggestion system based on an electronic health record phenotype may increase the uptake of clinical care pathways in the care workflow. Such real-time identification of patients with acute gastrointestinal bleeding is a crucial first step in developing an EHR-based model for risk prognostication. By mapping out the points at which data is generated from the process model, this guides the eventual deployment and implementation of a risk stratification prognostic algorithm for clinical decision making in real time.

Development and Validation of Long-Short-Term-Memory Recurrent Neural Networks for Dynamic Risk Prediction of Red Blood Cell Transfusion in Patients with Acute Gastrointestinal Bleeding Requiring Intensive Care Unit Stay

Introduction

Acute gastrointestinal bleeding accounts for over 2.2 million hospital days and 19.2 billion dollars of medical charges annually in the United States and frequently requires red-blood cell transfusion.[119] The management of severe acute gastrointestinal bleeding begins with resuscitation using intravenous fluids and transfusion of packed red blood cells, which are given to 43% of patients hospitalized with upper gastrointestinal bleeding in the United Kingdom and 21% of patients hospitalized with lower gastrointestinal bleeding in the United States.[120,121]

Transfusion needs may change during the hospital stay, but a tool to dynamically predict transfusion needs over time does not yet exist in clinical care. Patients with severe acute gastrointestinal bleeding who require care in the intensive care setting generally have higher transfusion needs and may benefit most from a predictive tool to guide resuscitation efforts. Current guidelines are based on a restrictive transfusion strategy using a hemoglobin threshold of 7g per deciliter compared to the previous threshold of 9g per deciliter in patients with upper gastrointestinal bleeding.[122]

Dynamic risk prediction, where predictions are generated in real time every hour based on clinical and laboratory values, may help guide transfusion strategies and help in timing endoscopic intervention, particularly in severely ill patients who require intensive care. Existing clinical risk scores used to screen for risk of needing transfusion of packed red blood cells, such as the Glasgow-Blatchford Score, are static models that only use clinical information at the time of admission (e.g. initial systolic blood pressure).[123] Machine learning approaches to model risk for gastrointestinal bleeding have shown promise in outperforming existing clinical risk scores, but are also static models.[124,125] Electronic health records (EHRs) can capture clinical data in real time, and have been used to create automated tools to model adverse events, such as sepsis, post-operative complications, and acute kidney injury.[126-129] Recurrent neural networks, a type of neural network that accepts time series data and sequences, have been demonstrated to be better than state-of-the-art risk models for continuous prediction of acute kidney injury up to 48 hours, the onset of septic shock 28 hours before onset, and all-cause inpatient mortality.[130-132] We propose the use of a Long-Short-Term Memory (LSTM) Network, an advanced recurrent neural network, to process data from electronic health records with an internal memory that stores relevant information over time and can generate a probability of transfusion within the 4 hour intervals for patients with severe acute gastrointestinal bleeding. LSTMs have the advantage that feature modules carefully decide what information to store and what information to discard, thereby offering the potential for improved performance. Figure 15 shows the use of our LSTM model in an example patient with generated risk predictions throughout the first 24 hours from admission.

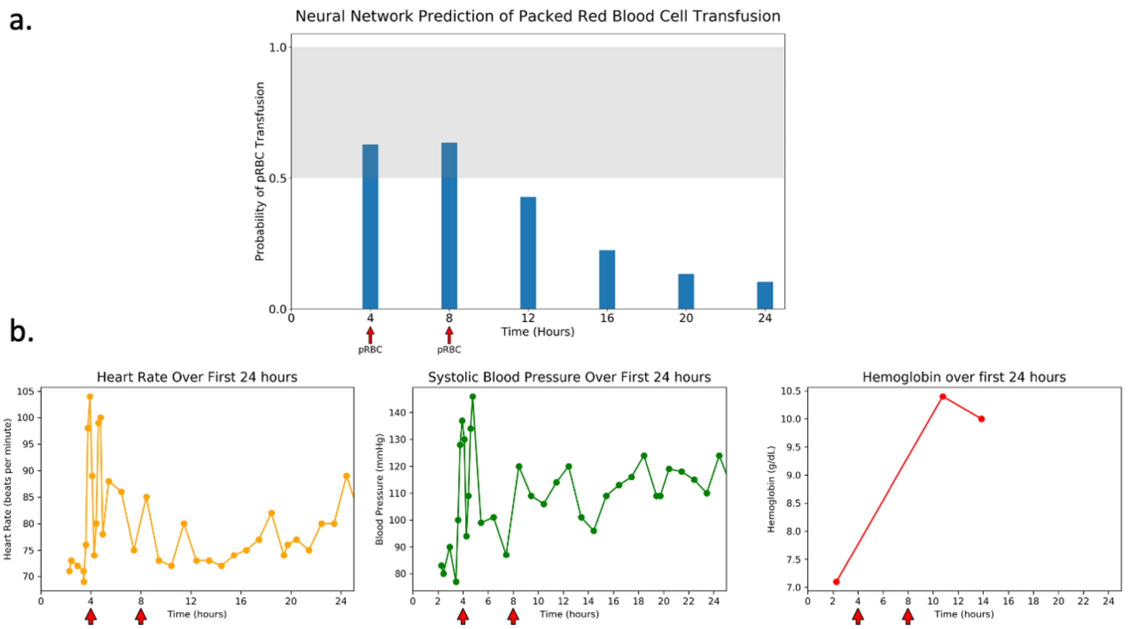


Figure 15: Example of neural network prediction for the first 24 hours of a 62-year-old man with Hepatitis C cirrhosis presenting with 2 days of intermittent coffee ground emesis and lethargy. Initial Glasgow Blatchford Score = 14 a) Continuous risk prediction of the neural network through the first 24 hours with the threshold set above 0.5 for detecting need for transfusion. The arrows indicate need for transfusion during that time period. b) Measurements of Heart Rate, Systolic Blood Pressure, and Hemoglobin occurring during the first 24 hours.

Methods

Data Source

A patient cohort presenting with acute gastrointestinal bleeding was identified from the Medical Information Mart for Intensive Care III (MIMIC-III) critical care database.[133,134] The database contains data for over 40,000 patients in the Beth Israel Deaconess Medical Center from 2001 to 2012 requiring an ICU stay. For external validation, a patient cohort presenting with acute gastrointestinal bleeding was extracted from the Phillips eICU Collaborative Research Database (eICU-CRD) of critical care units across the United States from 2014 to 2015. Only urban hospitals with greater than 500 beds were included.

Patients were included if they had an admission diagnosis containing the terms “gastro”, “bleed”, “melena”, “hematochezia”. The diagnoses were collated and then manually reviewed. This inclusion criteria were meant to specifically capture patients with severe acute gastrointestinal bleeding requiring ICU stay. Patients were excluded if vital signs were only available greater than 24 hours from time of admission to the ICU, since this constitutes missing values for all 4-hour time intervals used to train the models. The data included information that was updated over time during the course of hospitalization, including laboratory results and vital signs. For laboratory values, any negative entry or non-quantizable (e.g., >=, <) was converted to missing. Medications, current procedural terminology codes, and ICD9/10 codes from the visit were excluded from the analysis. The dataset had a total of 62 features: 5 clinical and demographic variables and 57 laboratory variables.

Data Access

All clinical data from MIMIC-III was approved under the oversight of the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA). The Phillips eICU Collaborative Research Database (eICU-CRD) was under the oversight of the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived by both institutional review boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of

Technology because the project did not impact clinical care and all protected health information was deidentified. All procedures were performed in accordance with relevant guidelines. The data was available on PhysioNet were derived from protected health information that has been de-identified and not subject to HIPAA Privacy Rule restrictions. All use of the data was performed with credentialed access under the oversight of the data use agreement through PhysioNet and the Massachusetts Institute of Technology.

Study Design

The MIMIC cohort included 2,524 hospital admissions and was randomly split into a training set with 2,032 hospital admissions and an internal validation set with 492 hospital admissions. (Table 17)

Table 17: Demographics and Baseline Data for the Training and Validation Set

	Training Set N = 2,032		Validation Set N = 492			External Validation Set N = 1526		
	N	Prop	N	Prop	p-value	N	Prop	p-value
Demographic Information								
Male	836	41%	190	39%	0.31	919	59%	<0.01
Age								
>89	144	7%	42	9%	0.29	57	4%	<0.01
75-89	629	31%	168	34%	0.24	438	28%	0.06
50-75	935	46%	200	41%	0.14	808	52%	<0.01
25-50	316	16%	71	14%	0.43	211	14%	<0.01
<25	8	0%	4	1%	0.27	12	1%	0.13
Ethnicity								
White	1429	70%	380	77%	0.08	1246	79%	<0.01
African American	244	12%	52	11%	0.35	172	11%	0.35
Hispanic	75	4%	22	4%	0.37	27	2%	<0.01
Asian American	74	4%	15	3%	0.42	20	1%	<0.01
Other	210	10%	23	5%	0.05	54	3%	<0.01
Clinical Features								
Upper Gastrointestinal Bleeding	679	33%	203	41%	0.07	666	43%	<0.01
Lower Gastrointestinal Bleeding	428	21%	162	33%	0.02	448	29%	<0.01
Unspecified Location	925	46%	127	26%	<0.01	412	27%	<0.01
Outcomes								
Packed Red Blood Cells	1542	76%	381	77%	0.39	515	33%	<0.01
In-Hospital Mortality	156	8%	32	6.5%	0.35	103	6.6%	0.21
	Mean	Std Dev	Mean	Std Dev	p-value	Mean	Std Dev	p-value
Vital Signs								
Heart Rate (beats per minute)	88.9	18	88.1	16.6	0.35	86.8	17.8	<0.01
Systolic Blood Pressure	126.9	22.9	127.1	22.2	0.86	119	23.3	<0.01
Diastolic Blood Pressure	64.2	16.9	65.7	16.6	0.07	61.7	15.4	<0.01

We chose to compare the model to a logistic regression model, a standard approach to prediction for time-varying electronic health record data that has previously been applied to acute kidney injury.[135] We also compared the model to a regularized regression model, which uses additional parameters to optimize prediction.[136] The eICU cohort included 1,526 hospital admissions from 12 large urban hospitals with over 500 beds. The performance of the neural network model and the regression-based models were compared on the internal validation dataset and the external validation dataset.

Input Variables

A total of 62 input variables were used and included age, gender, vital signs (systolic blood pressure, diastolic blood pressure, heart rate), and 57 unique laboratory values. (Table 18)

Table 18: Input Variables (N = 62)

Category	Input Variables
Demographic (2)	Gender Age
Vital Signs (3)	Heart Rate Systolic Blood Pressure Diastolic Blood Pressure
Laboratory Variables (57)	Blood Gas (Base Excess, Total Carbon Dioxide, Oxygen Saturation, pH, Arterial Pressure of Oxygen) White Blood Cells, Neutrophils, Basophils, Eosinophils, Lymphocytes, Bands, Monocytes, Hemoglobin, Hematocrit, Mean Corpuscular Hemoglobin, Mean Corpuscular Hemoglobin Concentration, Mean Corpuscular Volume, Red Blood Cell Distribution Width, Platelet Count, International Normalized Ratio, Prothrombin Time, Partial Thromboplastin Time Sodium, Potassium, Chloride, Bicarbonate, Anion Gap, Magnesium, Phosphate, Calcium, Creatinine, Urea Nitrogen, Glucose Alanine Aminotransferase, Aspartate Aminotransferase, Alkaline Phosphatase, Albumin, Amylase, Lipase, Direct Bilirubin, Total Bilirubin Creatine Kinase, Creatine Kinase-MB, Ferritin, Total Iron, Iron Binding Capacity, Lactate, Lactate Dehydrogenase, Thyroid Stimulating Hormone, Transferrin, Troponin T, Vancomycin, Fibrinogen Urine Studies (Creatinine, Sodium, Specific Gravity)

The vital signs and laboratory values were extracted and then consolidated into 4-hour time intervals over the first 24 hours from admission. These features were selected because they reflect dynamic changes from measurement in the ICU; ICD codes and CPT codes associated with the encounters were not included since they are not available at the time of care provision and therefore not available in real time for prediction. Medications have different formulations, with no clear definition of relevant medication types or standardization across multiple centers and were not included as features for this analysis.

Outcome Variable

The predicted outcome measure was the transfusion of packed red blood cells, calculated as binary 0 (no transfusion) or 1 (transfusion given). At the beginning of each 4-hour time interval, the model makes a prediction on whether a transfusion will be needed at the next 4-hour interval.

Data Pre-Processing

Each patient encounter was represented by a sequence of events with each 4-hour period containing information recorded in the vitals and laboratory values. Information for each patient encounter was encoded into 4-hour time intervals up to the first 24 hours. After excluding lab values with greater than 90% missingness, remaining lab values with greater than 50% missingness in the dataset were converted to missing indicator variables, with 1 as present and 0 as missing. To harmonize the input variables across patients, the first timepoint for each patient encounter was fixed at the first recording of heart rate, systolic blood pressure, and diastolic blood pressure. Consolidation of vital signs and laboratory values in each 4-hour interval was performed by taking the mean of each value. All continuous values were normalized and centered. Age was maintained as a continuous variable, with patients greater than 89 years old coded as 89 years old. After consolidation, 86% (1651/1923) of the encounters had information for every 4-hour interval in the full 24-hour period. For the training set 7% of the 4-hour periods (855/13167) were labeled as receiving a packed red blood cell transfusion, the test set 4% (134/3149), and the external validation set 2% (157/8414). In summary, each patient encounter has up to 6 predictions for a total of 6*n predictions in the entire dataset, and we compute one ROC curve and associated AUC for this total. This ensures that the same threshold exists across every time period.

Missing Values

To examine the role of the data imputation method used, we compared 4 different imputation strategies. The first was imputation of the mean value for any missing value. The second was a carryforward approach, or using the previously recorded value if a value was present at a previous time point but no subsequent value was measured. This assumes that the laboratory value is constant until the next time point in clinical decision-making.[137] The third was mean imputation with a new variable that served as a missingness indicator for every variable. The fourth was carryforward with a missingness indicator for every variable.

LSTM Neural Network Model background

Recurrent neural networks allow for processing of sequential information by storing information as internal states over multiple time points. Long short-term memory (LSTM) networks are a type of RNN that can be useful for clinical measurements because they carefully tune the information passed between subsequent time-iterations of the model (Figure 16).

T represents the time in hours, X represents input data (vitals, laboratory values), Y represents the probability of needing transfusion, and FCN is a fully convolutional network that processes the information from the previous time period to generate the prediction.

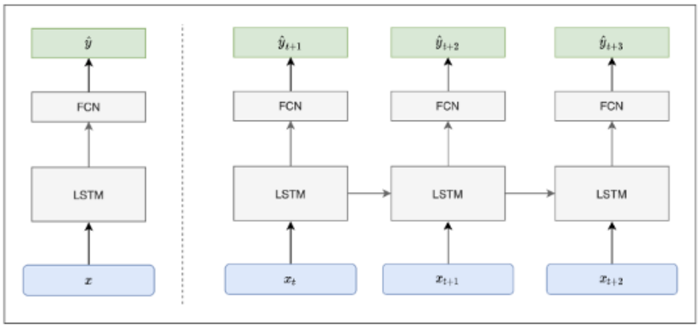


Figure 16: Long-Short Term Memory (LSTM) Network Model Overview. Electronic Health Record data (vitals, laboratory values) is fed into the model, passed through the layers, transformed, and gives a probability of the outcome (transfusion of packed red blood cells). At the beginning of each 4-hour interval the LSTM Network can generate a probability of needing transfusion.

The LSTM has a single output that serves as a prediction and other hidden states that are then fed back into the neural network to adjust the final output. For the implementation of the model, we used the PyTorch deep learning library. Given a series of EHR data, $\boldsymbol{x}^{(\boldsymbol{0})}$, $\boldsymbol{x}^{(\boldsymbol{1})}, \ldots,$ $\boldsymbol{x}^{(\boldsymbol{T-1})}$, where $\boldsymbol{x}^{(\boldsymbol{t})}$ represents the input variables for the $(t+1)$th 4-hour interval, at the beginning of each 4-hour interval our goal is to predict whether transfusion is needed in the next 4 hours. The output is a sequence of probability predictions $\widehat{y^{(1)}}, \widehat{y^{(2)}}, \ldots, \widehat{y^{T}}$, where $\widehat{y^{(t)}} \in [0,1]$ is the prediction for whether transfusion is needed in the t^{th} 4-hour interval. The LSTM model consists of 2 layers of 128 LSTM cells each, followed with a linear layer that maps from hidden state space to the prediction space. We obtain the log-probabilities by adding a LogSoftmax later in the last layer of the network. Thus, the output of the neural network is a sequence $\widehat{\boldsymbol{p}^{(1)}}, \widehat{\boldsymbol{p}^{(2)}}, \ldots, \widehat{\boldsymbol{p}^{(T)}}$, where $\widehat{\boldsymbol{p}^{(t)}}$ is the log-probability of $\hat{y}$ being either of the target classes, and our decision rule is to administer transfusion if $\widehat{\boldsymbol{p}^{(t)}} > threshold$, where the threshold is determined by desired sensitivity or specificity. We use the negative log likelihood for the output at each time of interest as the loss function. The model is trained for up to 100 epochs with hyperparameters corresponding to the lowest validation loss recorded and used to obtain testing accuracy.

Discrete Time Logistic Regression and Regularized Regression

For comparison discrete-time regression approaches were employed to generate a new prediction using each 4-hour block of data to predict the need for transfusion for the next 4-hour block of data. We used both logistic regression and regularized regression with elastic net penalty using the glmnet package in R tuned by fivefold cross-validation on the training set (Figure 17).

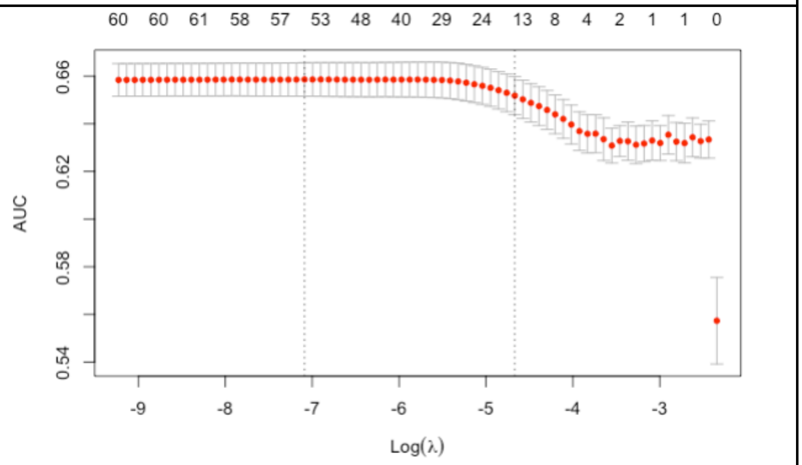
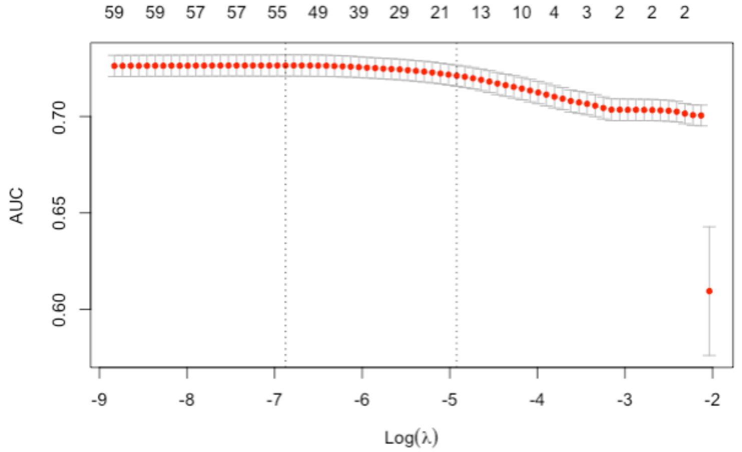
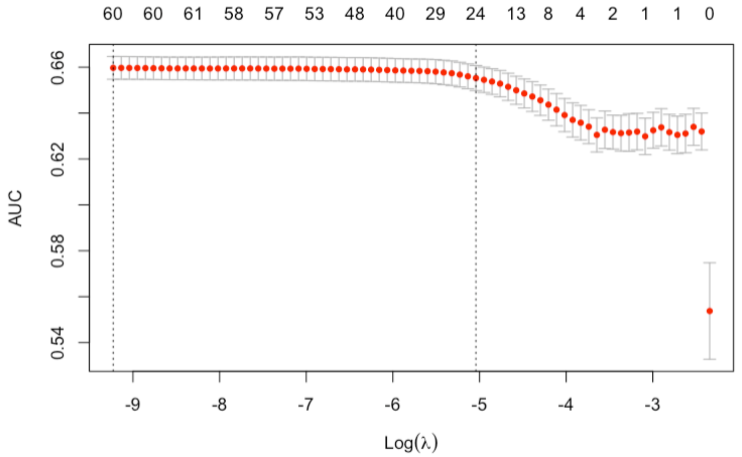
Imputation Strategy	**Lambda Selection with Cross Validation**
Mean Imputation	
Carryforward Imputation	
Mean Imputation and Missing Indicators	

Figure 17: Comparison of Imputation Strategies for Lambda Selection with Cross Validation.

The training protocol was to take every 4-hour sequence and then using all the 4-hour sequences to train the regression models, since the model is designed to generate a prediction for any 4-hour sequence. The same covariates were used that were available for the LSTM neural network model at each 4-hour time interval, with no additional features used to train the model. The different imputation strategies as described previously were also employed.

Statistical Analysis

Two-tailed t tests and chi-squared test were used to compare baseline characteristics between the training and validation sets. We assessed model performance using the area under the curve (AUROC) and compared it to the performance of logistic regression using the nonparametric DeLong test.[138] Confidence intervals were calculated with 2000 stratified bootstrap replicates. McNemar's test was used to compare the optimal sensitivity and specificity threshold by the Youden Index.

Results

Demographics were similar between training and internal validation sets with the median age 69 for both, proportion of men (41% in training, 39% on internal validation), and predominantly white (70% in training, 77% in internal validation). There was a similar percentage of patients with upper gastrointestinal bleeding (training 33% vs internal validation 41%), but the training set had more patients with gastrointestinal bleeding from an unspecified source (46% vs 26% *P*<0.01), while the internal validation set had more patients with lower gastrointestinal bleeding (33% vs 21% *P*=0.02). Vital signs and laboratory values were similar in the training and internal validation sets. (Table 17) The external validation set was significantly different from the training and internal validation with demographics notable for a generally younger population, increased patients with upper and lower gastrointestinal bleeding and less patients with an unidentified source. Furthermore, the transfusion rate was significantly lower (33% versus 76%;*P*<0.01), reflecting modern guidelines of restrictive transfusion strategy for the treatment of acute gastrointestinal bleeding. Laboratory tests were notable for decreased hemoglobin and hematocrit, increased ALT, AST, alkaline phosphatase and total bilirubin, increased creatinine and decreased albumin. (Table 17).

The performance of the LSTM model on the four different imputation strategies were similar and all significantly better than the discrete time logistic regression model. (Table 18) The results we subsequently present are for the strategy with the highest AUROC (carryforward and missing indicators).

Table 18: Comparison of the overall performance of Long-Short Term Memory network model compared to the Logistic Regression model with different imputation methods to address missingness in the first 24 hours after admission for all patients admitted to the Intensive Care Unit with Acute Gastrointestinal Bleeding.

External Validation Set	LSTM AUROC 95% CI	Logistic Regression AUROC 95% CI	p-value	Regularized Logistic Regression with Elastic Net AUROC 95% CI	p-value
Mean Imputation	0.65 (0.60-0.69)	0.54 (0.49-0.59)	<0.001	0.55 (0.50-0.60)	<0.001
Carryforward Imputation	0.66 (0.62-0.70)	0.56 (0.51-0.60)	<0.001	0.56 (0.51-0.60)	<0.001
Mean Imputation and Missing Indicators	0.64 (0.60-0.68)	0.54 (0.49-0.59)	<0.001	0.55 (0.50-0.60)	<0.001
Carryforward Imputation and Missing Indicators	0.65 (0.60-0.69)	0.56 (0.51-0.60)	<0.001	0.56 (0.52-0.61)	<0.001

For the main analysis of all patients with acute gastrointestinal bleeding who were transferred to the ICU, the LSTM performed significantly better than both regression-based approaches. On internal validation, the LSTM outperformed LR (AUROC 0.81 CI 0.80-0.83 vs 0.75 CI 0.73-0.77;*P*<0.001) and regularized regression (AUROC 0.81 CI 0.80-0.83 vs 0.75 CI 0.73-0.78;*P*<0.001) in predicting packed red blood cell transfusion across the entire 24-hour period. For external validation, the LSTM outperformed LR (AUROC 0.65 CI 0.61-0.69 vs 0.56 0.51-0.60;*P*<0.001) and regularized regression (AUROC 0.65 CI 0.61-0.69 vs 0.56 0.52-0.61;*P*<0.001). (Table 19, Figure 18)

Table 19: Performance of the Long-Short Term Memory (LSTM) Model and the discrete time Logistic Regression (LR) model in Predicting Transfusion of Packed Red Blood Cells by

Comparison of Area Under the Receiver Operating Curve (AUROC) for Internal Validation (N = 492) and External Validation (N=1526).

	Long-Short Term Memory Network Model AUROC 95% Confidence Interval	Logistic Regression AUROC 95% Confidence Interval	p-value	Regularized Logistic Regression with Elastic Net AUROC 95% CI	p-value
Internal Validation	0.81 (0.80-0.83)	0.75 (0.73-0.77)	<0.001	0.75 (0.73-0.78)	<0.001
External Validation	0.65 (0.61-0.69)	0.56 (0.51-0.60)	<0.001	0.56 (0.52-0.61)	<0.001

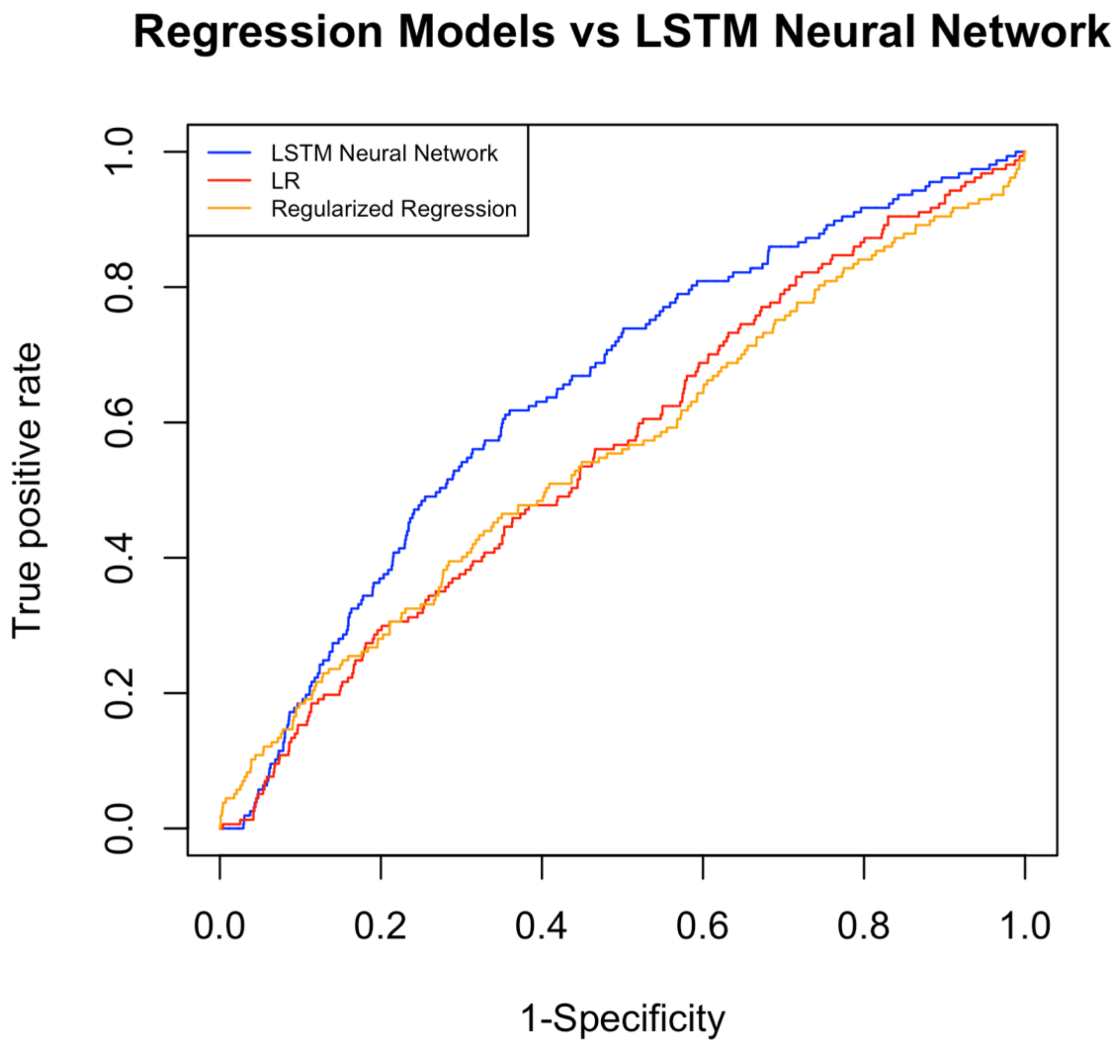


Figure 18: Comparison on external validation of the overall Area Under the Receiver Operating Curve (AUROC) as a measure of performance of the Long-Short Term Memory (LSTM) Neural Network model and discrete time Logistic Regression (LR).

Sensitivity and Specificity Cutoff

The optimal sensitivity and specificity cutoff was obtained using Youden's index and was found on external validation for the LSTM neural network to be 62% sensitivity and 64% specificity; the logistic regression optimal cutoff was 47% sensitivity and 65% specificity (P<0.001).

Sensitivity Analysis

Sensitivity analysis was performed on the external validation dataset by gender, age, systolic blood pressure, blood urea nitrogen, and hemoglobin, variables commonly used in assessing risk for patients with acute gastrointestinal bleeding. When subset by gender the LSTM model still outperformed the LR model (0.64 vs 0.54, P=0.002) and the regularized regression model (0.64 vs 0.49;*P*=0.02). In the subset of patients with age greater than 65, which was the mean of patients with acute gastrointestinal bleeding, the LSTM model outperformed the LR model (0.61 vs 0.54, P=0.008) and the regularized regression model (0.61 vs 0.56;*P*=0.01). For vital signs and laboratory values, cutoffs were derived from the Glasgow Blatchford Score: systolic blood pressure cutoff less than 110 mmHg, blood urea nitrogen greater than 18.2, and hemoglobin less than 10 (similar risk category for both men and women). The LSTM model outperformed the LR and regularized regression models in all these analyses. (Table 20) Sensitivity analyses of the opposite group (men only, patients less than 65 years old, and the lower risk cutoff for vital signs and laboratory values) are provided. (Table 20)

Table 20: Sensitivity Analyses for external validation eICU dataset. Systolic Blood Pressure, BUN, and hemoglobin cutoffs were derived from the Glasgow Blatchford Score. Hemoglobin cutoff was chosen due to the matched risk for both men and women.

Total Encounters N=1526	LSTM Neural Network	Logistic Regression	p-value	Regularized Logistic Regression with Elastic Net	p-value
Female N=607	0.64 (0.57-0.71)	0.54 (0.46-0.62)	0.002	0.49 (0.42-0.56)	0.02
Age >65 (mean) N=820	0.61 (0.55-0.67)	0.54 (0.47-0.60)	0.008	0.56 (0.50-0.61)	0.01
Systolic Blood Pressure <110 N=849	0.64 (0.58-0.69)	0.57 (0.50-0.63)	0.03	0.55 (0.49-0.61)	0.002
BUN >18.2 N=1110	0.64 (0.58-0.69)	0.56 (0.50-0.62)	0.003	0.56 (0.50-0.61)	<0.001
Hemoglobin <10 N=1303	0.64 (0.60-0.69)	0.56 (0.51-0.62)	0.001	0.56 (0.51-0.61)	<0.001

Discussion

Predicting the need for transfusion of packed red blood cells has direct relevance to guiding the management of patients with acute gastrointestinal bleeding. This is the first study to show that a LSTM network model is able to predict the need for packed red blood cell transfusion for patients with severe acute gastrointestinal bleeding with superior performance to time-varying logistical regression with internal and external validation. By anticipating needs for transfusion, this is a first step towards personalizing treatment and tailoring appropriate resuscitation to reduce clinical decompensation and death for patients with severe acute gastrointestinal bleeding. While endoscopic evaluation is important, adequate resuscitation is an important part of management prior to endoscopy.[9,139-141]

In this work we use a (one-directional) 2-layer LSTM with 128 hidden units in each layer. The LSTM setup is a commonly used variation of the LSTM which consists of the original LSTM architecture with added forget gates and full gradient backpropagation through time (BPTT) training.[142,143] We use this model over a simple recurrent neural network (SRNN) as it addresses weaknesses inherent in SRNNs such as difficulty learning dependencies across multiple time steps and aberrant gradient flow. A comparative study of LSTM variants concluded that while many variations of LSTMs exist, much of the improved performance can be attributed to forget gates and the choice of activation function.[144] Advantages of the LSTM over regression models include the ability to generate multiple predictions with the first data input and the ability to combine features in more complex ways to model changes over time. The trained architecture can be used to generate predictions for each time period using presenting data from the first 4 hours, whereas the regression models have fixed coefficients that can only generate predictions as data becomes available for each time period. For example, for a patient admitted to the ICU with data from the first 4 hours, the LSTM neural network can propagate the data through its architecture to predict need for transfusion at 8, 12, 16, 20, and 24 hours. Using regression models, it could only be used to predict the need for transfusion at the next time period. While regression models weighted sums of features are used with specific thresholds for prediction, neural networks can combine features in non-linear and more complex ways to generate predictions.

Previous risk scores capture information from specific points in time at admission, and do not incorporate new clinical data over the course of hospitalization. Electronic health records contain longitudinal information on patients admitted to the hospital and reflect real-world practice, which can be used to develop risk prediction models.[145] For patients who have severe disease requiring intensive care unit stay, mortality may be more due to end organ damage due to inadequate perfusion.[121,146,147] Despite the significant computing requirements necessary to run neural networks, existing electronic health records are now deploying cloud computing infrastructure able to perform computationally intensive tasks. The emerging capabilities of cloud infrastructure in electronic health records, such as the Cognitive Computing platform for Epic Systems, make the deployment of neural networks for clinical care feasible.

We envision the future of care for all patients to be enhanced by customized machine learning decision support tools that will provide both initial risk stratification and ongoing risk assessment to provide treatment at the right time for the right patient. Using a dynamic risk assessment, resuscitation needs could be estimated early and optimized in preparation for endoscopic evaluation and intervention. This individualized decision-making potentially will minimize organ damage from inadequate resuscitation, which drives the risk for mortality in these patients.[146] The LSTM model can be tuned for provider preference. Alert fatigue is particularly relevant in the ICU, since clinically irrelevant alerts can have an impact on patient safety.[148] In order to minimize alert fatigue, a high specificity threshold could be set for the algorithm. However, if providers do not want to miss any time periods when patients need packed red blood cell transfusions, a high sensitivity threshold can be set to minimize false negatives. Although the LSTM network model is much better than a standard regression-based approach, it still falls short of optimal performance. More work will be needed to develop and validate neural network models.

Interpretability is a key area of active research for neural network models, particularly in order to assess the trustworthiness of the prediction. Approaches attempt to elucidate the hidden states of the network architecture, identify features important to prediction, and perform saliency analyses to identify input data most relevant to the model prediction.[149-152] Another approach attempts to learn an interpretable model around the prediction, called Local Interpretable Model-agnostic Explanations (LIME).[153] These approaches, however, should be filtered through the usefulness for a front-line clinician who has both prior knowledge about the application and the ability to reason through the available evidence after receiving the prediction. As professionals with authority due to training and experience, clinicians may benefit less from the "hidden states" and more from presenting the relative importance of input variables; the latter allows for clinicians to assess the prediction as plausible or due to confounding.[154] Applying these techniques is outside the scope of this manuscript and will be explored in future work.

Strengths of this study include external validation in a more recent ICU electronic health record dataset and modeling patients with severe illness requiring intensive care unit stay, which may benefit disproportionately from timely transfusion and resuscitation and the use of vital signs and laboratory tests that are standardized and can be easily mapped across electronic health record systems. Our comparison to regression models is stronger than currently used clinical scores such as the Glasgow-Blatchford Score or Oakland Score, which were developed to generate a static risk prediction with only data at presentation.

Limitations include the absence of prospective and independent validation in other electronic health record-base datasets. Despite showing external validation on a temporally and geographically separate dataset of patients with acute gastrointestinal bleeding requiring ICU care, prospective validation and implementation into clinical practice is crucial to quantifying the benefit of such systems on patient outcomes. Additionally, the performance difference between test set and validation set may be due to the lower prevalence of packed red blood cell transfusions in the external validation set, which may indicate need for re-training of the model with more updated clinical data that reflect the decreased use of transfusions. The definition of ground truth is the receipt of a transfusion, and not on the judgment of whether they should have received a transfusion, which may not reflect the current standard of care and may not be applicable to hospitals that are resource limited. The use of encounters as independent episodes rather than individual patients may lead to bias and information leak, particularly since there are around 708 patients with more than one encounter for severe acute gastrointestinal bleeding requiring ICU care. However, the decision was made to include all encounters for these patients to reflect real world practice since the bias is tolerable from a clinical standpoint: patients with recurrent severe acute gastrointestinal bleeding requiring ICU care are the very patients who would stand to benefit from these predictions. We also control for information leak since all features except for age and sex and unique for each ICU encounter. Comparison with regression-based models may change if the models incorporate aggregated data available at time of predictions from previous time intervals (e.g., the mean and standard deviation) and should be explored in future studies. In addition, the segmentation into 4-hour segments may lead to distortions, since the same signal of transfusion can be administered immediately after bound of the 4-hour time interval or several hours afterwards (e.g., 5 minutes or 2 hours afterwards). Additionally, the proportion of missing data required imputation, which may introduce bias to the data. To quantify the difference, we compared different imputation strategies including carryforward and found no clear difference in the overall performance of the models.

In summary, we present the first application of recurrent neural networks to dynamically predict need for packed red blood cell transfusion over time using electronic health record data. We report superior performance compared to a discrete time regression models. Our approach may lead to delivery of earlier resuscitation with packed red blood cells to minimize ischemic end organ damage in patients with severe acute gastrointestinal bleeding. Future directions include external validation of the model on other cohorts of high-risk patients with gastrointestinal bleeding, along with prospective implementation and deployment in the electronic health record system for high-risk patients with gastrointestinal bleeding.

Modeling Missingness in Clinical Data with Variables Missing-Not-At-Random with MURAL: An Unsupervised Random Forest-Based Embedding for Electronic Health Record Data

Introduction

Unsupervised nonlinear embedding methods have allowed for exploration manifold learning of big high dimensional datasets in many fields ranging from epidemiology, to biology, to physics. However, a major limitation of using unsupervised embeddings in healthcare data is the large amount of missingness in the data as well as the mixed modality of the variables collected. In a typical EHR or patient dataset the range of missing data range from 20% to 80%, varying across broad categories of possible fields such as demographics, laboratory values, and treatment information.[155-157] Further there is a mix of real-valued, categorical and binary data which can be difficult to normalize or scale. This makes it difficult to compute distances and affinities between datapoints—the first step in nonlinear dimensionality reduction methods such as tSNE[158], UMAP[159], diffusion maps[160] or PHATE[161]. Similar distance/affinity computations are also required for spectral clustering[162], which operates on a graph Laplacian computed from the affinity matrix. Thus, data with missing values cannot be used, and if the values are MNAR they cannot be imputed.

To tackle these issues, we propose to use an intermediary representation called the MURAL-forest, an unsupervised random forest in which tree distances between datapoints form an accurate measure of dissimilarity and can be used for data distance/affinity computation, as needed in methods specified above.[159-161,163]. MURAL creates a set of trees by splitting on any variable type (categorical, continuous, with or without missingness) using a marginal entropy criterion that is computed on *other* variables. Further, MURAL ensures that heterogeneity within categorical or MNAR variables is immediately broken down using low dimensional entropy to create 4-way splits at such levels. We test MURAL on ground truth data that the resulting tree distances result in accurate embeddings.

While random forests are normally supervised and trained for prediction, there have been some efforts to learn random forests in an unsupervised manner.[164] describes a method called *manifold forests* which effectively use a splitting criterion based on intra-versus-inter split affinity or density.

However, these and other methods often presuppose the ability to compute distances or affinities between high dimensional datapoints. By contrast, we use our MURAL unsupervised random forests in order *to be able to compute* an accurate distance between datapoints with missing and mixed-mode variables.

We show the accuracy of our method by comparing the MURAL derived distances to known ground truth and recovering embeddings in a 5-dimensional Swiss roll. We then apply our method to a complete case subset of an intensive care unit dataset and of an international patient registry dataset of patients presenting with symptoms of upper gastrointestinal bleeding. We induce missingness in the complete case subsets in specific ranges of laboratory values and compare imputed values using mean imputation and multiple imputation with chained equations to the original ground truth. We then construct MURAL-embeddings on the full datasets with missingness. We show that MURAL-embeddings consistently display more structure and create separations that are more clinically meaningful than commonly used imputation methods. Finally, we show an application of our method in comparing entire cohorts of patients by computing a tree-based Wasserstein distance on the MURAL-forest, which can be used to quantify similarities or distances between patient cohorts.

Background

A. Manifold Learning, Dimensionality Reduction, Clustering

Though there are many nonlinear dimensionality reduction and embedding methods, we focus our results on methods that can learn the *data manifold* or intrinsic low dimensional shape and structure of the data. We believe that this is useful in biomedical settings where many measurements of the patient reflect non-orthogonal aspects of the same underlying entity, essentially indicating the data in fact lies in a lower dimensional space.

High dimensional data can often be modeled as a sampling $Z = z_{i}{}_{i=1}^{N} \subset M^{d}$ of a *d* dimensional manifold M^{d} that is mapped to *n* dimensional observations $X = x_1, \dots, x_N \subset \mathrm{R}^{\mathrm{n}}$ via a nonlinear function $x_i = f(z_i)$. Intuitively, although measurement strategies, modeled here via *f*, create high dimensional observations, the intrinsic dimensionality, or degrees of freedom within the data, is relatively low. This manifold assumption is at the core of the vast field of manifold learning[160,163,165,166], which leverages the intrinsic geometry of data, as modeled by a manifold, for exploring and understanding patterns, trends, and structure that displays significant nonlinearity.

Diffusion maps were proposed as a robust way to capture intrinsic manifold geometry in data by eigendecomposing a powered diffusion operator. Using t-step random walks that aggregate local affinity is able to reveal nonlinear relations in data and allow their embedding in low dimensional coordinates. These local affinities are commonly constructed using a Gaussian kernel:

$$\mathbf{K}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right), \quad i, j = 1, ..., N \quad (1)$$

where K forms an $N \times N$ Gram matrix whose (i,j) entry is denoted by $K(x_i,x_j)$. A diffusion operator is defined as the row-stochastic matrix $P = D^{-1}K$ where D is a diagonal matrix with $D(x_i,x_i) = {}^{P}{}_{j} K(x_i,x_j)$. The matrix P, or diffusion operator, defines single-step transition probabilities for a time-homogeneous diffusion process, or a Markovian random walk, over the data. Furthermore, powers of this matrix P^t, for $t > 0$, can be used to simulate multi-step random walks over the data, helping understand multiscale organization of X, which can be interpreted geometrically when the manifold assumption is satisfied. P has been used in many downstream unsupervised learning tasks, eigendecomposition of P yields the popular diffusion map dimensionality reduction method that can be used as input to clustering. P is also used by the PHATE method for visualization. PHATE transforms the diffusion operator with a pointwise logarithm log(P), derives distances between points x_i,x_j as $k\log(P_i) - \log(P_j)k_2$, and then embeds the resulting distances, known as *potential distances*, with metric MDS.

Other methods for visualization such as tSNE and UMAP use K rather than P to focus on near neighbors rather than learning the entire data manifold. The diffusion operator P is related to the graph Laplacian that, depending on the normalization used, can be written as $L = I - K$ or $L = I - P$. Thus the graph Laplacian has the same eigenvectors and eigenvalues that are in the opposite order. Spectral clustering is often described in terms of the graph Laplacian, i.e., *k*-means over a graph Laplacian rather than data.

B. Decision Trees

A *tree T* is a connected directed acyclic graph $T = (V,E)$ with vertices (or nodes) $V = t_1, t_2, \ldots, t_n$ and $n-1$ edges E such that every node has at most one incoming edge. A rooted tree has a *root* node t_1 with no incoming edges, while t_i, $i > 1$, all have exactly one incoming edge. A node $t_j \in$ *children*(t_i) if and only if $[t_i, t_j] \in E$, i.e., there is a directed edge from t_i to t_j. A *descendant*(t_i) is any node t_k that is connected to t_i by a directed path $t_i, \ldots, t_k$ width a directed edge between each consecutive pair of nodes.

Decision trees contain nodes that split on a variable to create partitions of the data such that datapoints on one side of the partition are more similar to each other in terms of the decision variable. Recursive splits create finer granularity branches where data points are similar with respect to all of the variables that have been split on the path to the node. A specific strength of decision tree is the ability to naturally split multiple types of data—binary, ordinal, and missing.

C. Supervised Random Forests

In classification tasks single decision trees can learn irregular patterns and overfit to data. As a way of addressing this, random forests average over sets of decision trees[167] and are created by randomizing variable splits. The algorithm selects a random subset of features at each potential split, and chooses a threshold so as to optimize a local criterion such as the *Gini impurity index* or *information gain*. The Gini impurity index is an information theoretic measure that is based on Tsallis entropy.[168] For *C* classes (given labels) with fractions $P = p_1, p_2, \ldots, p_C$ of observations in each class, the Gini impurity index is given by $I_G(P) = 1 - \sum_i p_i$. Information gain is also an information theoretic measure which measures the difference in Shannon entropy between the parent node and child nodes. Shannon entropy of a probability distribution *P* is given by $H(P) = -\sum_i p_i \log(p_i)$. Information gain is defined as

$$I_G(P) = H(P) - \sum_a \frac{|a|}{k} H(P^a). \qquad (2)$$

Here *P* is the class distribution of the parent node, and P^a is the class distribution of the *a*-th child node, which receives |*a*| datapoints. The total number of datapoints split by the parent node is *k*. Note that these criteria are with respect to a classification label that is given in a supervised setting.

The original random forest classifier used labeled data to randomly train an ensemble of decision trees with a majority vote aggregating the classifications. Decision trees are constructed through recursively partitioning the space occupied by data as observations travel from the tree's root to its leaves, each nonterminal node containing a weak learner that chooses a splitting variable and threshold. These weak learners minimize an impurity function to ensure that each child node receives a "purer" cohort than its parent. Purity is determined by the proportion of labels; if all examples belong to the same class, the subset is considered pure.

D. Unsupervised Random Forests

Variants of decision trees have been used to cluster data in the absence of labels: random projection trees[169,170], density forests[164], PCA trees[171], approximate principal direction trees[172], and geodesic forests.[173] These variants are often effective at learning the manifold of the data when the data variables are continuous and distances or Gaussian affinities can be defined between datapoints. However, for us this creates a chicken-and-egg problem. Our purpose in creating a random forest is to derive a meaningful distance in situations where there are missing values and categorical variables, where simple Euclidean distances are not meaningful.

For example, Criminisi's manifold forests use trees whose nodes minimize the following information gain measure when splitting

$$I_G(S_j) = \log(|\Lambda(S_j)|) - \sum_{i \in \{L,R\}} \frac{|S_j^i|}{|S_j|} \log(|\Lambda(S_j^i)|). \qquad (3)$$

Here, S_j is the set of datapoints that node j partitions, S_j^L and S_j^R are the sets of datapoints from S_j that get sent to the left and right child of node j, respectively. The matrix $\Lambda(S)$ is a set's covariance matrix, which is undefined in our case with missing values. Furthermore, unless binary affinities are chosen, the affinity matrices defined using manifold forests depend on preexisting distances between datapoints. Thus we define a new type of tree that can tolerate missing values and mixtures of variables, which can itself be used to compute a new type of distance.

E. Wasserstein Distance over Trees

The 1-Wasserstein distance (also known as the earth mover's distance) measures the total cost of moving shifting the mass from one probability distribution to another. For discrete probability distributions over a general metric space this can be computed exactly in $O(n^3)$ time using the Hungarian algorithm[174], and approximated using entropic regularization in $O(n^2)$ time.[175] However, for discrete probability distributions over a tree metric space the 1-Wasserstein distance can be computed exactly in linear time.[176] Given two probability distributions μ, v over a measurable space Ω with metric $d(\cdot,\cdot)$, let $\Pi(\mu,v)$ be the set of joint probability distributions π on the space $\Omega \times \Omega$, where for any subset $\omega \subset \Omega$, $\pi(\omega \times \Omega) = \mu(\omega)$ and $\pi(\Omega \times \omega) = v(\omega)$. The 1-Wasserstein between μ and v is defined as:

$$W_\rho(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} \rho(x,y)\pi(dx,dy). \qquad (4)$$

Let k · k$_{L\rho}$ denote the Lipschitz norm w.r.t. ρ, when Ω is separable w.r.t. ρ and μ, v have bounded support, then the dual of equation 4, known as the Kantorovich-Rubinstein dual, can be expressed as:

$$W_\rho(\mu,\nu) = \sup_{\|f\|_{L_\rho} \le 1} \int_\Omega f(x)\mu(dx) - \int_\Omega f(y)\nu(dy). \qquad (5)$$

When d is a tree metric over a rooted tree T, for every pair of points $x,y \in \Omega$, $\rho(x,y)$ is the total weight of the (unique) path between nodes x and y in T. Denote the edge weight associated with each node *t* as w_t, and *D*(*t,μ*) as the sum of mass of *μ* at and below node *t*, then the Wasserstein distance between two distributions on *T* can be expressed as:

$$W_{\rho_T}(\mu,\nu) = \sum_{t \in T} w_t \left| D(t,\mu) - D(t,\nu) \right| . \qquad (6)$$

Previous work in demonstrated unsupervised forest constructions that approximate the Wasserstein distance when *ρ* is the Euclidean ground metric over Ω ≡ R^d. In MURAL, we construct an unsupervised random forest over a high dimensional Ω that consists of continuous, categorical and missing variables. These trees subsequently define a distance on Ω, which in turn defines a Wasserstein distance between distributions on Ω, and because of the specific construction of MURAL trees, admits a simple feature importance measure described in IV-C.

III. MURAL

Next, we present the MURAL algorithm for building unsupervised random forests from continuous and categorical data with missing values on healthcare data. Our code is available at https://github.com/KrishnaswamyLab/MURAL.

A. Problem Formulation

Our goal is to build a distance matrix *D* whose (*i,j*)-th entry contains the distance $d(x_i,x_j)$ between observations x_i and x_j. Desirable properties for *D* are that neighbors found using *D* have similar clinical manifestations, and moreover that *D* can be used in a nonlinear dimensionality reduction method to create nonlinear axes corresponding to largest patient variation, and clusters that group patients by overall similarities.

B. Distinguishing Randomly versus Non-Randomly Missing Variables

A key insight in MURAL is that healthcare data consists of variables that are intentionally missing, i.e., missing not at random (MNAR), that are a source of significant information since the data is related to unobserved patient characteristics (e.g. there appears to be higher levels of missing values for reported income in individuals with higher income levels). However, patient data also often contains data that is missing completely at random (MCAR) or missing at random (MAR). MCAR data is when missingness does not depend on the observed or missing values, and MAR data is when missingness does not depend on the missing values but may depend on the observed values. MCAR or MAR data are usually the result of absence of documentation through the extraction, transformation and loading of clinical data.[177,178] We note that MCAR or MAR variables can be imputed on the basis of informational redundancy with other variables using conditional probability modeling, regression or other techniques, with multiple imputation leading to the most unbiased results.[179,180] However, MNAR variables cannot be imputed well as we show in Section IV.

MURAL distinguishes between MCAR variables and MNAR features for continuous variables. We may or may not have prior knowledge about which continuous variables are MCAR vs. MNAR. If we do not, then we can distinguish between the two cases based on Little's test[181], which examines patterns of missingness for correlations with other variables. Little's test gives each variable a significance value for rejecting the null hypothesis that says it is missing at random. If this significance value $p \leq 0.05$ then we conclude that the variable is MNAR, otherwise we deem the variable to be MCAR. Next, for variables that are MNAR we use imputation to fill the values in as a preprocessing step using fully conditional specification (FCS) multi-variable imputation.[180] After this step, all variables are either fully imputed or MNAR.

C. Mixed-modality Variable Splitting Scheme

MURAL incorporates multimodal variables into an unsupervised random forest framework by using a nuanced splitting scheme. Key aspects of the MURAL splitting scheme are as follows:

- At each iteration MURAL chooses a variable v_i to split on at random.
- If the v_i is MNAR for some observations then we create a preliminary binary split between observations where it is measured and observations where it is missing.

- For the branch where the variable is measured, we find a single threshold based on the unsupervised information gain criterion we define in the following section to create two child nodes.

- For the branch where the variable is missing, we randomly select another variable v_j with no missingness and create two child nodes based using the *residual multidimensional entropy* described in Equation 7.

- If v_i has no missingness then it is split into two child nodes again using the same information gain criterion.

- The two-level splits described above are flattened into the same level to create four child nodes if the first split was missing/not-missing (see Figure 19).

This scheme effectively creates four child nodes when a variable has MNAR since there are actually two variables worth of information, first variable describing the missingness, and second variable describing a further split. However, the reason for choosing an additional variable on the branch with missingness is to create a branch with controlled heterogeneity (entropy) instead of the forced split that comes with a naturally binary variable.

A similar four-way splitting scheme can avoid fragmentation caused by the presence of binary variables. For a binary variable v_b, we create two hidden nodes, one for the value of 0 and one for the value of 1, and choose a second, non-binary, variable v_c for the hidden nodes to optimize binary splits on. The edges are weighted as in the case of missingness, resulting in a four-way split.

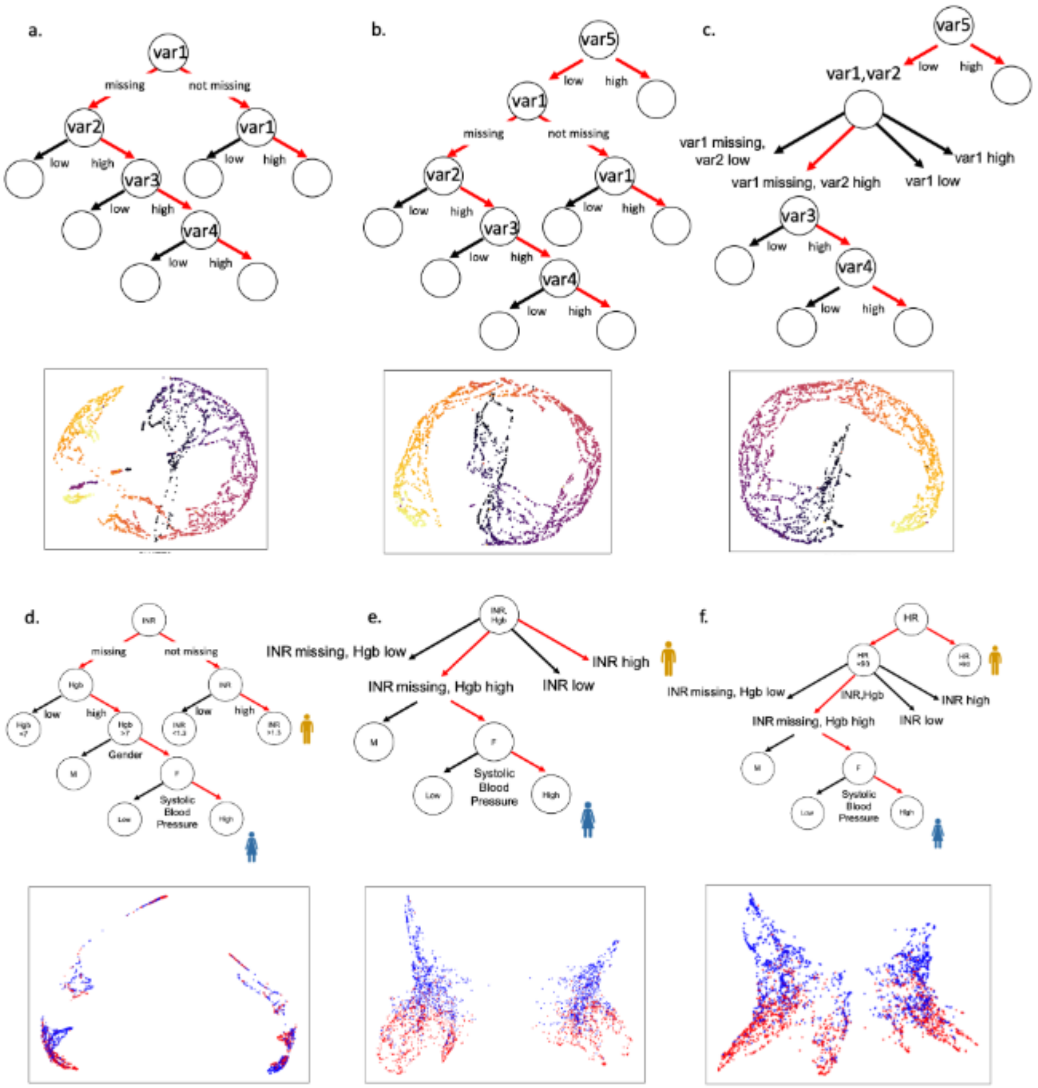


Figure 19. The decision to create the four-way split for variables missing not at random (MNAR) and to avoid splitting on MNAR variables at the root were based on empirical findings as shown above. For the 5-dimensional Swiss roll, in a) no four-way split and no condition to avoid splitting MNAR variables at the root node results in fragmentation. In b) addition of condition to avoid splitting MNAR variables at the root decreases fragmentation and c) introduction of four-way splitting results in clearer structure recovery. For the upper gastrointestinal bleeding dataset, in d) no four-way split and no-condition to avoid splitting MNAR variables at the root node results in

distorted structure e) addition of four-way splitting allows for structure recovery, and f) introduction of condition to avoid splitting MNAR variables at the root leads to clearer recovery

D. Splitting with Residual Multidimensional Entropy

When each variable without MNAR is chosen, we choose a threshold *thres*(v_i) based on an unsupervised information gain criterion, which we term *residual multidimensional entropy*. Instead of choosing a split that maximizes the information gain in the class label, we choose a split that maximizes the residual entropy of the remaining variables. Since the remaining variables are not naturally discrete like class labels, we discretize the continuous variables. Thus for each node t_i splitting on variable v_i we compute the following summation:

$$I_G(S_i) = H(S_i) - \sum_{a \in \{L,R\}} \frac{|S_i^a|}{|S_i|} H(S_i^a) \tag{7}$$

Where S_i is the probability distribution of the classes (discretizations) of variable v_i (not on the path from the root node to v_i) among *descendants*(t_i), and S_i^a is the probability distribution of the classes of v_i among only the descendants of *child*(t_i,*a*) i.e. the *a*th child node. We discretize each variable into a number of bins which are determined by the Sturges method [30] (number of bins is dlog$_2$ *n*e + 1, where n is the number of datapoints).

Here, we avoid using a high dimensional entropy or high dimensional density for ease of computation. We assume that we can approximate this entropy by a sum of marginal entropies or by the multidimensional entropy of a subset of the variables. In our experiments, we found that using three dimensional entropies calculated over randomly chosen subsets of variables resulted in the best embeddings of the Swiss roll dataset. For datasets with more dimensions (such as UGIB and eICU), three variables would cover only a small fraction of the information available, so we prefer the sum of marginal entropies. As noted before, we cannot directly use within-split affinity or density estimates as that is indeed the end result of MURAL.

E. MURAL-derived Distances and Embeddings

For every tree T_k in a trained forest, we define the tree distance between two nodes to be

$$d_k(t_i, t_j) = \sum_{p \in \mathcal{P}} w_{t_p} \qquad (8)$$

where P is the index sequence of the edges in T_k on the unique undirected path (t_i,...,t_j) between t_i and t_j with no node repetition, and w_{tp} is the weight of edge *p*. We opt for all edges having unit weights.

Then we define distances between x_i,x_j by noting that nodes corresponding to leaves of the tree $t_{l,1}$,$t_{l,2}$,... each contain sets of datapoints, i.e., $t_{l,1} \ = \ x_{l,1}, x_{l,2} \ldots x_{l,m}$. Thus if *xi* ∈ *tl,i* and *xj* ∈ *tl,j* then *dk*(*xi,xj*) = *dk*(*tl,i,tl,j*)

The constructed tree metrics can be averaged over the MURAL-forest, which results in composite MURAL distances.

$$D_M = (1/k) \sum_k D_k. \qquad (9)$$

Here D_k is the matrix of pairwise distances of the *k*th tree, and D_M the averaged distance matrix over all trees.

This leads directly to a MURAL-based distance matrix, which can be converted into an affinity matrix using any kernel function. For embedding, the distance matrix can be passed as an input to a nonlinear dimensionality reduction algorithm such as PHATE [7], which we choose for its manifold affinity preservation capabilities. We call the resultant PHATE embedding using D_M the *MURAL-embedding*.

Results

A. Empirical Validation

In this section we induced missingness, i.e., MNAR, in datasets (or data subsets) where there was no missingness to validate the ability of MURAL to recover meaningful distances. We used 3 datasets for this experiment:

1) A synthetic Swiss roll constructed of 3,000 points embedded in a 5-dimensional space.

2) A dataset of patients requiring intensive care unit stay (ICU) from the publicly available Phillips eICU Collaborative Research Database (eICU-CRD) of critical care units across the United States from 2014 to 2015.[182,183] Patients with available data within the first 24 hours of ICU stay (148,532 unique patient encounters) had 10 variables selected: five laboratory values (bilirubin, blood urea nitrogen, creatinine, hematocrit, and albumin), age, 3 ordinal variables from nursing assessment of eye, motor, and verbal responses using the Glasgow Coma Scale, and binary variable of invasive mechanical ventilation within the first 24 hours.

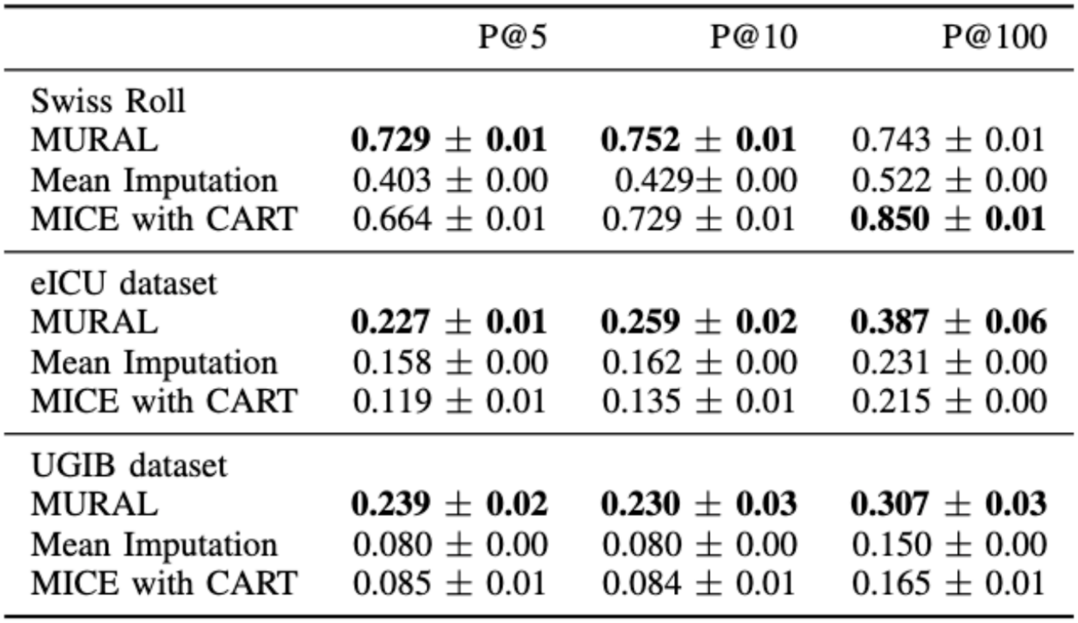
3) An international registry of consecutive, unselected patients presenting with symptoms of upper gastrointestinal bleeding between March 2014 and March 2015 from Yale–New Haven Hospital (United States), Glasgow Royal Infirmary (Scotland), Royal Cornwall Hospital Truro (England), Odense University Hospital (Denmark), Singapore General Hospital (Singapore), and Dunedin Hospital (New Zealand).[184] 7 variables were selected: 4 laboratory variables (hemoglobin, urea, albumin, INR), 1 vital sign (systolic blood pressure), 1 binary demographic variable (gender), and 1 ordinal variable (degree of liver disease).

The eICU and UGIB datasets already have missingness in a significant portion of the entries. Thus, to create an artificial ground truth, we used only entries with all variables present. For the UGIB dataset, 2,761 patients with complete data were selected. For computational efficiency we subsampled the eICU dataset to 10,000 patient encounters with complete data across 10 variables. Using this subset of complete data as ground truth, we artificially induced missingness in order to test the ability of MURAL to recover meaningful distances.

We induced missingness in the Swiss roll dataset in a similar way to what is observed in real clinical data, with one variable that has a pattern of missing values deemed to be missing not at random by pairwise Little's test, and two variables with random values dropped out. In the eICU dataset we induced missingness by dropping the values of bilirubin > 3, the threshold chosen since these are clearly physiologically abnormal values and thus would be missing not at random. In the UGIB dataset we induced missingness by dropping the values of INR > 3, also chosen since they are physiologically abnormal values and would be missing not at random. All datasets were standardized after missingness was induced but before they were used for constructing MURAL-forests.

To quantitatively compare the preservation of the underlying manifold structure of these datasets in the presence of missing values, we used the accuracy of a kNN graph derived from the embedding compared to the ground truth kNN graph from each of these graphs. We compared the performance to mean imputation, and to another standard tree-based imputation method, multiple imputation with chained equations (MICE) using classification and regression decision trees. We find that near neighbors are recovered in the MURAL-embedding with greater accuracy than baselines that first use imputation in this case where ground truth is known in IV-A. In the artificial Swiss roll dataset, the MURAL-forest with 100 trees outperformed mean imputation at 5, 10, and 100 neighbors by 32%, 30%, and 22%, and outperformed MICE with CART at 5 and 10 neighbors by 6% and 3%. This reflects the specific nature of the Swiss roll dataset, since smaller neighborhoods were more likely to be perturbed due to the intrinsic coiled data structure manifold. For the eICU dataset and the UGIB dataset, the MURAL-forest with 100 trees outperformed mean imputation at 5, 10, and 100 neighbors by 7%, 10%, and 15%; MURAL-forest outperformed MICE with CART at 5, 10, and 100 neighbors by 15%, 15%, and 14%. (Table 21)

Table 21. $\mu \pm \sigma$ for P@5, 10, 100 metrics on three datasets over 5 runs. MURAL-embedding preserves neighborhoods for missing values better than mean imputation and MICE with CART.



	P@5	P@10	P@100
Swiss Roll			
MURAL	**0.729 ± 0.01**	**0.752 ± 0.01**	0.743 ± 0.01
Mean Imputation	0.403 ± 0.00	0.429± 0.00	0.522 ± 0.00
MICE with CART	0.664 ± 0.01	0.729 ± 0.01	**0.850 ± 0.01**
eICU dataset			
MURAL	**0.227 ± 0.01**	**0.259 ± 0.02**	**0.387 ± 0.06**
Mean Imputation	0.158 ± 0.00	0.162 ± 0.00	0.231 ± 0.00
MICE with CART	0.119 ± 0.01	0.135 ± 0.01	0.215 ± 0.00
UGIB dataset			
MURAL	**0.239 ± 0.02**	**0.230 ± 0.03**	**0.307 ± 0.03**
Mean Imputation	0.080 ± 0.00	0.080 ± 0.00	0.150 ± 0.00
MICE with CART	0.085 ± 0.01	0.084 ± 0.01	0.165 ± 0.01

B. Ablation Study

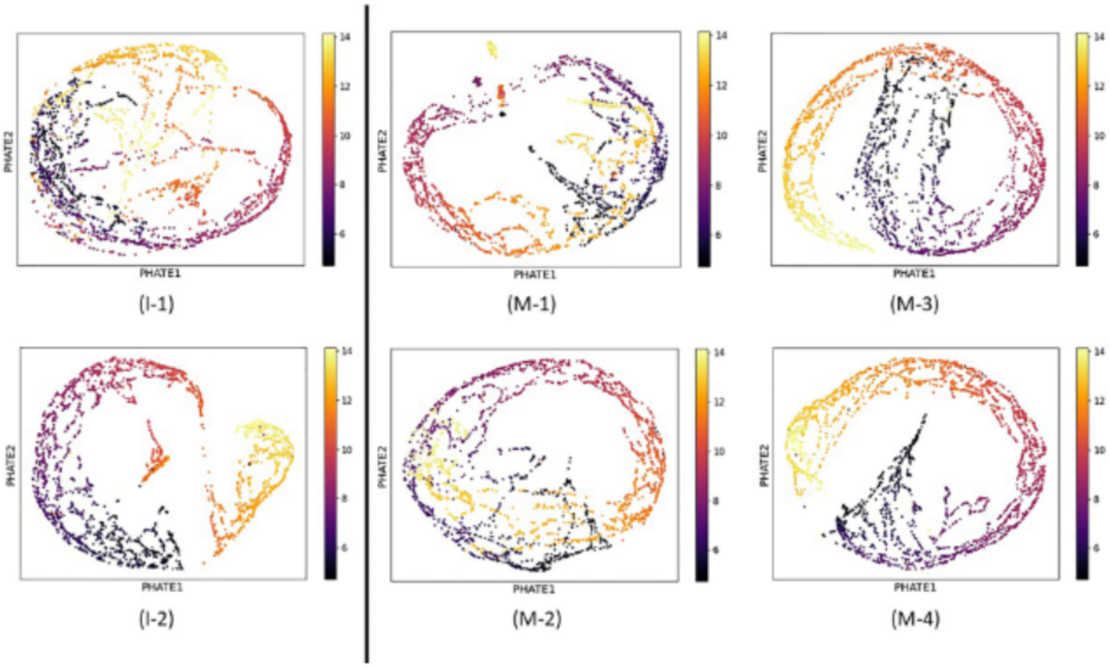
Next, we carried out an ablation study using this Swiss roll dataset to investigate which parameter choices lead to the best embeddings. Results are shown in Table 22. For each parameter choice, we trained MURAL-forests with 5 different random initializations. In the run testing 100 trees of depth 10, we rejected one outlier forest that generated pathological distances. Generally we find that low-dimensional (3d) entropy works for our proposed residual multidimensional entropy from Equation 7. Surprisingly we found that splitting on a *single* residual variables works best for the discrete and MNAR case. An additional insight was that restricting MNAR variables to low levels in the tree worked best as they would have minimal effect on other distances in this scheme. Full results are seen in Table 2.

Table 22. Ablation Study for MURAL-Embedding on the swiss roll embedded in 5 dimensions ($\mu \pm \sigma$) over 5 runs. BOLD represents best in each parameter.

Model	DeMAP	P@5	P@10	P@100	P@500	Distortion	Time (s)
Mean Imputation	0.495	0.637	0.635	0.581	0.675	5530	—
Entropy Dimensions							
1D	0.504 ± 0.08	0.720 ± 0.01	0.742 ± 0.01	0.739 ± 0.01	**0.783 ± 0.01**	1880 ± 270	376 ± 240
2D	0.618 ± 0.04	0.728 ± 0.01	0.748 ± 0.01	0.746 ± 0.01	0.782 ± 0.00	1800 ± 180	**150 ± 5.4**
3D	0.577 ± 0.08	0.733 ± 0.01	0.755 ± 0.01	**0.749 ± 0.01**	0.779 ± 0.01	1860 ± 250	240 ± 22
5D	**0.628 ± 0.03**	0.704 ± 0.01	0.725 ± 0.00	0.727 ± 0.02	0.763 ± 0.01	$1.23 \pm 2.0 \times 10^7$	1150 ± 260
RME	0.556 ± 0.06	**0.738 ± 0.01**	**0.757 ± 0.01**	0.738 ± 0.01	0.773 ± 0.01	**1540 ± 120**	448 ± 41
# Variables Split							
1	**0.582 ± 0.08**	0.733 ± 0.00	**0.755 ± 0.01**	**0.749 ± 0.01**	**0.779 ± 0.01**	**1860 ± 250**	**286 ± 23**
2	0.542 ± 0.07	0.752 ± 0.01	0.781 ± 0.01	0.725 ± 0.01	0.751 ± 0.01	2150 ± 81	517 ± 5.7
3	0.550 ± 0.05	0.757 ± 0.01	0.782 ± 0.01	0.721 ± 0.01	0.741 ± 0.01	2020 ± 200	989 ± 110
4	0.530 ± 0.06	**0.762 ± 0.00**	0.781 ± 0.00	0.711 ± 0.01	0.738 ± 0.01	2140 ± 180	1240 ± 140
Restrict MNAR Levels							
0	0.552 ± 0.05	0.712 ± 0.01	0.725 ± 0.01	0.573 ± 0.02	0.538 ± 0.03	2180 ± 360	274 ± 33
1	0.604 ± 0.04	0.722 ± 0.01	0.728 ± 0.00	0.703 ± 0.00	0.744 ± 0.01	2060 ± 230	297 ± 22
2	**0.636 ± 0.03**	0.716 ± 0.01	0.730 ± 0.01	0.737 ± 0.01	0.772 ± 0.00	1920 ± 170	304 ± 16
3	0.570 ± 0.08	**0.733 ± 0.00**	**0.755 ± 0.01**	**0.749 ± 0.01**	**0.779 ± 0.01**	**1860 ± 250**	**264 ± 34**
Tree Depth							
2	0.189 ± 0.01	0.477 ± 0.02	0.527 ± 0.02	0.609 ± 0.01	0.721 ± 0.02	$8.61 \pm 3.5 \times 10^8$	**55.7 ± 1.6**
4	0.259 ± 0.02	0.655 ± 0.01	0.688 ± 0.01	0.685 ± 0.01	0.769 ± 0.01	$7.17 \pm 1.3 \times 10^7$	110 ± 8.7
6	0.614 ± 0.05	0.707 ± 0.01	0.744 ± 0.01	0.722 ± 0.01	0.780 ± 0.01	$5.27 \pm 4.9 \times 10^6$	164 ± 10
8	0.623 ± 0.05	0.724 ± 0.01	0.746 ± 0.01	**0.747 ± 0.01**	**0.784 ± 0.01**	1950 ± 130	194 ± 16
10	0.624 ± 0.03	**0.733 ± 0.01**	**0.753 ± 0.01**	0.743 ± 0.01	0.776 ± 0.01	**1790 ± 67**	174 ± 26
12	0.623 ± 0.01	0.729 ± 0.01	0.750 ± 0.01	0.744 ± 0.01	0.773 ± 0.01	1940 ± 110	195 ± 1.3
14	**0.646 ± 0.03**	0.729 ± 0.01	0.751 ± 0.01	0.745 ± 0.01	0.774 ± 0.01	1833 ± 180	200 ± 5.8
Forest Size							
10	0.368 ± 0.13	0.597 ± 0.01	0.647 ± 0.01	0.702 ± 0.2	0.723 ± 0.02	$1.39 \pm 3.9 \times 10^8$	**41 ± 43**
50	0.596 ± 0.07	0.708 ± 0.01	0.734 ± 0.01	0.737 ± 0.01	0.765 ± 0.01	$5.94 \pm 8.0 \times 10^6$	108 ± 40
100	0.624 ± 0.03	0.733 ± 0.01	0.753 ± 0.01	0.743 ± 0.01	0.776 ± 0.01	**1790 ± 67**	174 ± 26
200	**0.632 ± 0.03**	0.746 ± 0.01	0.763 ± 0.01	0.755± 0.01	0.785 ± 0.01	1852 ± 120	332 ± 31
500	0.622 ± 0.03	**0.749 ± 0.01**	**0.769 ± 0.01**	**0.762 ± 0.00**	**0.791 ± 0.00**	1850 ± 61	801 ± 83

C. Applications

1) *Embeddings:* Our initial goal was to derive distances that provide a faithful representation of a dataset that contains different types of variable and missing values. In order to see if these distances facilitate embeddings that reveal structure and meaningful groupings in data, we fed the distance matrix into the PHATE nonlinear dimensionality reduction and visualization method. We chose to use PHATE due to its improved ability to preserve data manifold-affinities as quantified by the DeMAP metric. We note that PHATE contains similar information as a diffusion map, with information collected in low dimensions for visualization. The resultant embeddings, which we term as MURAL-embeddings are shown in Figure 20, Figure 21, Figure 22, Figure 23.



(I-1) (M-1) (M-3)

(I-2) (M-2) (M-4)

Figure 20. PHATE plots of the Swiss roll embedded in 5 dimensions. (I-1) Mean imputation. (I-2) MICE. (M-1) MURAL-embedding without any restrictions on choosing splits in variables with missing values. (M-2) MURAL-embedding with trees of depth 4. (M-3) MURAL-embedding with each node choosing the best split from among four variables. (M-4) MURAL-embedding with 100 trees of depth 10, each node choosing the best split from only one variable, not splitting on variables with missing values in the first three levels.

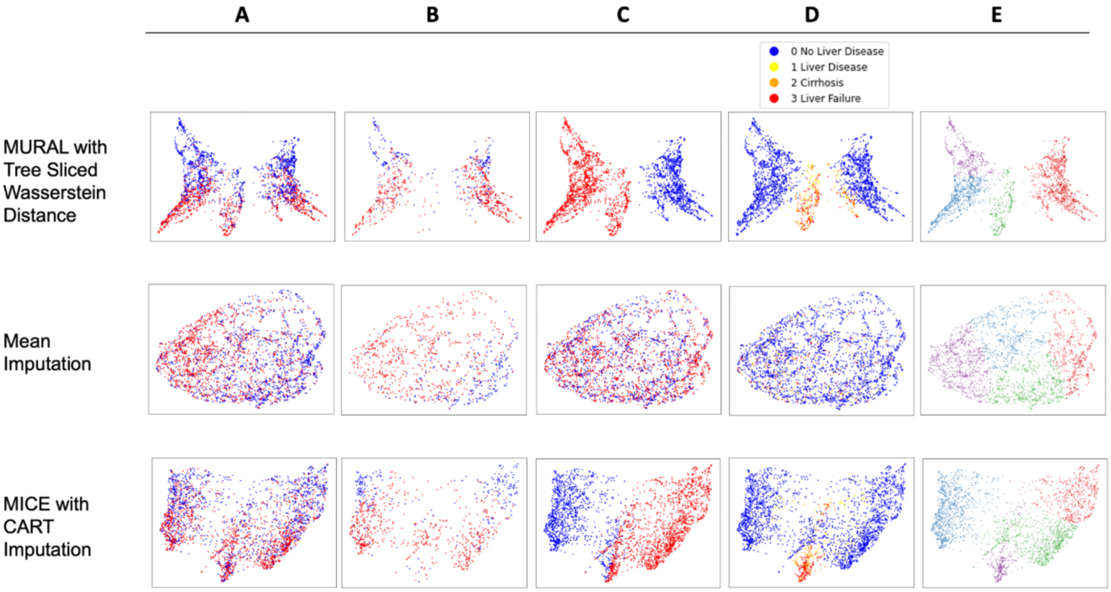


Figure 21. MURAL-embedding preserves data structure and separation better than mean imputation and multiple imputation with chained equations using classification and regression trees (MICE with CART) after visualization with PHATE. MURAL-embedding separates groups of patients with clinically relevant subgroups: A) high risk (red) and low risk (blue) groups as defined by need for hospital-based intervention B) different age groups <30 years old (blue) versus >80 years old (red) C) gender, male (red) versus female (blue) D) liver disease (yellow to red) versus no liver disease (blue). Spectral clustering (E) of the graph subsets the known groups including the group of men with liver disease (green), female patients (red), young males at low risk for hospital-based intervention (purple), and older males at high risk for hospital-based intervention (blue).

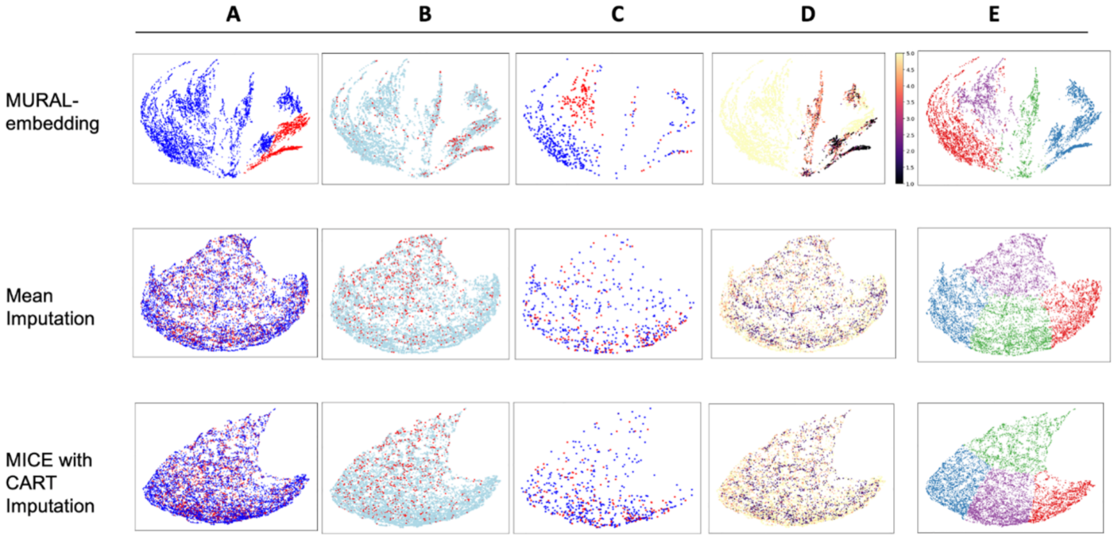


Figure 22. MURAL-embedding better separates out patients in the intensive care unit at risk for in-hospital mortality compared to mean imputation and MICE with CART on PHATE visualizations. MURAL-embedding separates patients who A) received invasive mechanical ventilation (red) versus not (blue) B) at high risk for in hospital mortality (red) versus not (light blue) C) presenting with admission diagnosis of acute coronary syndrome (red) versus other diagnosis (blue) D) nursing assessment (1 to 5, higher is healthier). Spectral clustering (E) of the graph subsets the groups of patients who required mechanical ventilation (blue), patients with slightly impaired verbal responses on nursing assessment (green), and patients with admission diagnosis of acute coronary syndrome with bilirubin measurement, reflecting concern about liver disease (purple) versus without bilirubin measurement (red), impairment of verbal responses (blue).

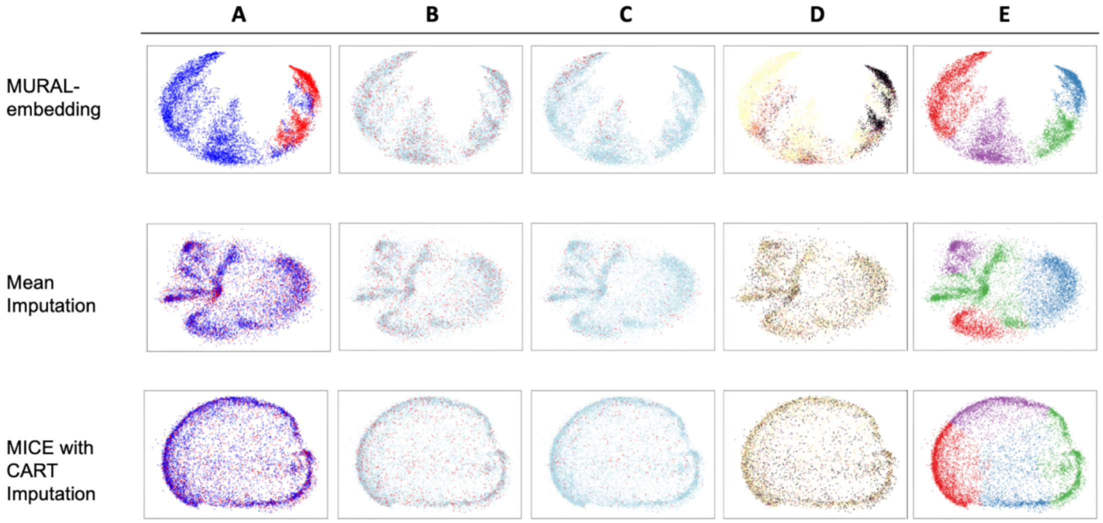
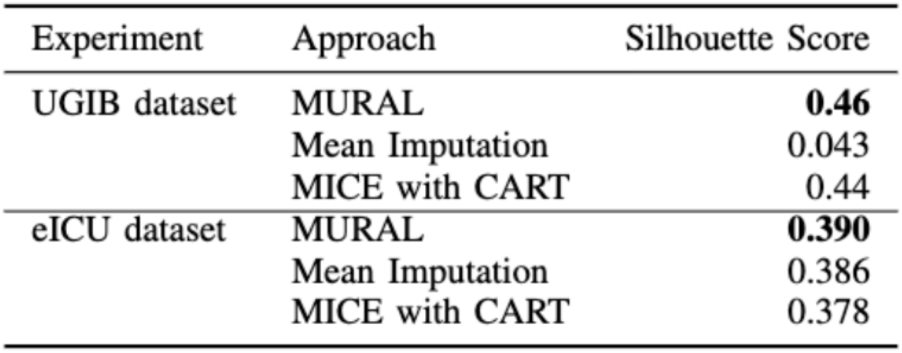


Figure 23. MURAL-embedding better separates out patients in the intensive care unit at risk for in-hospital mortality compared to mean imputation and MICE with CART on PHATE visualizations even with 108 variables. MURAL-embedding separates patients who A) received invasive mechanical ventilation (red) versus not (blue) B) at high risk for in hospital mortality (red) versus not (light blue) C) presenting with admission diagnosis of acute coronary syndrome (red) versus other diagnosis (blue) D) nursing assessment (1 to 5, higher is healthier). Spectral clustering (E) of the graph subsets the group of patients who required mechanical ventilation (blue/green).

By visual inspection we see that the MURAL-embeddings are much more structured than PHATE embeddings of the raw data with imputed values. Furthermore, visualization of several clinical variables on the embeddings show that the separations correspond to clinical groupings that are used in generating the embedding and also meaningful clinical groups not used to generate the MURAL-embeddings. In Figure 21 risk for hospital-based intervention (A) and age (B) were not used to generate the MURAL-embeddings, yet the MURAL-embeddings show separation into high risk and low risk groupings (A) and separation by age, in this case < 30 years old versus > 80 years old (B). For factors used to generate the MURAL embeddings, there are separations into two major structures horizontally based on gender (C) and minor structures based on liver disease status (D). By contrast, mean imputation mixes genders despite having the information, and MICE imputation overlaps the two disease cohorts. In Figure 22 risk for in-hospital mortality (B) and admission diagnosis of acute coronary syndrome (C) were not used to generate the MURAL-embeddings, but the MURAL-embeddings show separation into low and high-risk groups for in-hospital mortality (B) and subgroup of patients with admission diagnosis of acute coronary syndrome (C) (red) and those with other diagnoses (blue). Need for mechanical ventilation (A) and nursing assessment of verbal response (D) were used to generate the MURAL-embedding; clear separations are found between patients with mechanical ventilation (A, red) and those who did not (A, blue), as well as patients with normal verbal response (D, light yellow), mildly impaired verbal response (D, orange), those with no verbal response (D, purple). These separations are not seen in the embedding methods on imputed data despite also containing these variables. These visual results suggest that these embeddings are useful and amenable to further quantitative analysis. The structures in the MURAL-embedding are largely retained even after including 108 variables derived from the first 24 hours of patient stay in the ICU for the same patient cohort in the eICU dataset (Figure 23).

2) *Spectral Clustering:* Spectral clustering [8] is a key unsupervised clustering method that follows the data manifold by operating on a data affinity matrix. We used k-means with *k* = 4 on the diffusion operator created by PHATE, which is equivalent to spectral clustering and compared it to similar clustering using imputed data. To evaluate the qualities of the clusters, we use the silhouette score and find that clustered MURAL-embedding has highest silhouette score in both datasets. (Table 23)

Table 23. MURAL-Embedding on PHATE with spectral clusters K=4 shows superior silhouette scores compared to other imputation methods.



Experiment	Approach	Silhouette Score
UGIB dataset	MURAL	**0.46**
	Mean Imputation	0.043
	MICE with CART	0.44
eICU dataset	MURAL	**0.390**
	Mean Imputation	0.386
	MICE with CART	0.378

In addition, the resultant subgroups can be interpreted clinically. For example, the red cluster in Figure 21 corresponds to admission diagnosis of acute coronary syndrome who either have a bilirubin measured or not measured within 24 hours of ICU admission, which may suggest a concern for concurrent hepatobiliary dysfunction or injury. In Figure 21 the green cluster correspond to male patients with severe liver disease, which correspond to a specific type of gastrointestinal bleeding, portal hypertensive bleeding.

3) *Wasserstein Distance between cohorts:* Seeing that groups of patients who differ in a clinically significant way form distinct clusters on the embeddings, we decided to quantitatively check whether MURAL-forests themselves separate dissimilar cohorts more than similar ones. This is directly relevant to the task of characterizing clusters within the data representations. For example, if clusters are two groups of patients with the same diagnosis, if there is a meaningful difference between their measured laboratory or clinical characteristics that could reflect risk for a poor clinical outcome. To that end, we calculated tree-sliced Wasserstein distances between the low risk and high-risk patient cohorts. For the eICU dataset, we defined this as patients who died in the hospital, and in the UGIB dataset we defined this as a need for hospital-based intervention (red blood cell transfusion, hemostatic intervention, or 30-day mortality). We then extract feature importances by aggregating the variables associated with nodes where the Wasserstein distances were most disparate. The feature importances in Figure 24 are consistent with the top two variables predictive of in-hospital mortality from supervised machine learning algorithms (regression models) trained and validated on the eICU dataset[185], and the top three variables predictive of need for hospital based intervention from a high performing supervised machine learning algorithm (gradient boosted decision trees) trained and validated on the UGIB dataset.[14] As a sanity check, we also compared tree-sliced Wasserstein distances between different age groups: first, a very different age group (< 30 years old versus > 80 years old) and a similar age group (< 30 years old versus 30-40 years old). The similar age group had much lower TSWD compared to the group with different ages, and the TSWD from the MURAL-forest for high risk versus low risk was more distinct compared to the other imputation methods. (Table 24) More generally, we believe these types of

Wasserstein distances can be used to measure distances or similarities between treatments, diagnostic variants and other differences between sub-cohorts.

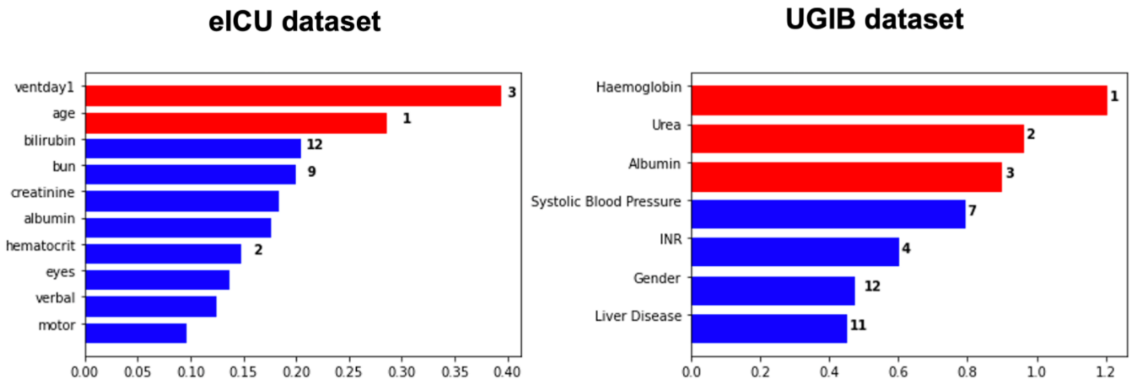
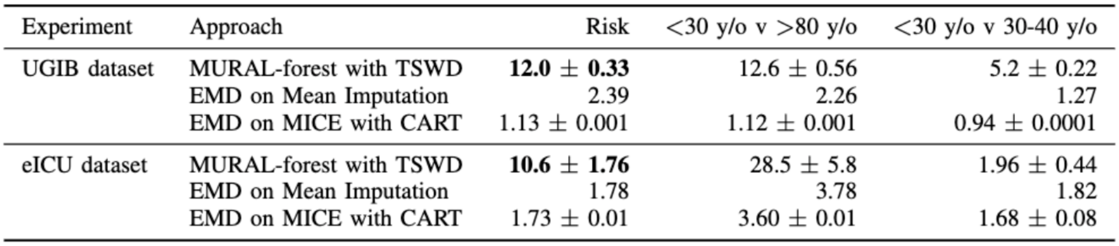


Figure 24. TSWD on the MURAL-forest can be used to generate feature importance graphs that are consistent with feature importances in supervised approaches on the same data. For the eICU dataset, the first two factors were within the top 3 predictive factors in supervised models predicting mortality on the same dataset. For the UGIB dataset, the first 3 factors were identified as the top 3 predictive factors in high-performing supervised models on the same dataset.

Table 24. Tree Sliced Wasserstein Distances (TSWD) on the MURAL-forest compared to mean imputation and multiple imputation with chained equations using classification and regression trees (MICE with CART). TSWD on the MURAL-Forest appear to separate populations by clinically meaningful risk (either in-hospital mortality for patients in the ICU or need for hospital-based intervention for patients with UGIB) more definitively than the other approaches, with an increased ratio of Earth Mover's Distance (EMD) of the two defined populations to EMD between random splits. Risk based on need for hospital-based intervention (UGIB) and in-hospital mortality (eICU).



Experiment	Approach	Risk	<30 y/o v >80 y/o	<30 y/o v 30-40 y/o
UGIB dataset	MURAL-forest with TSWD	**12.0 ± 0.33**	12.6 ± 0.56	5.2 ± 0.22
	EMD on Mean Imputation	2.39	2.26	1.27
	EMD on MICE with CART	1.13 ± 0.001	1.12 ± 0.001	0.94 ± 0.0001
eICU dataset	MURAL-forest with TSWD	**10.6 ± 1.76**	28.5 ± 5.8	1.96 ± 0.44
	EMD on Mean Imputation	1.78	3.78	1.82
	EMD on MICE with CART	1.73 ± 0.01	3.60 ± 0.01	1.68 ± 0.08

Semi-Supervised Classification Using MURAL-Forest

When utilizing the MURAL-forest trees in a semi-supervised task, the performance as measured by classification accuracy with kNN from randomly selected 10% to 90% random selection of labeled data. The performance is maintained with small neighborhoods to larger neighborhoods.

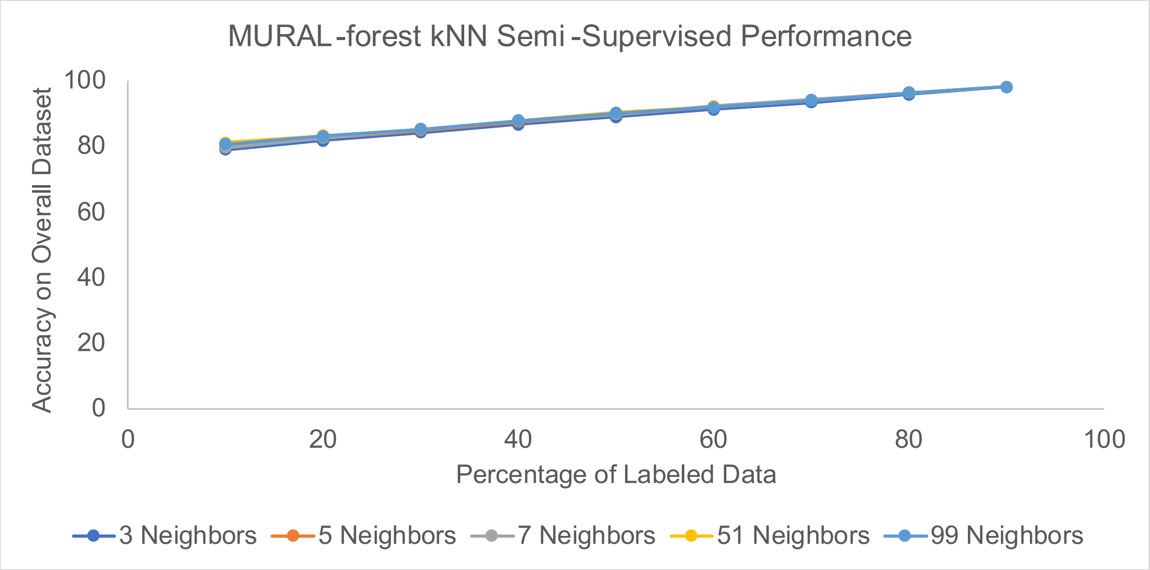


Figure 25. MURAL-forest kNN semi-supervised performance is maintained at small and larger neighborhoods across the MURAL-forest.

Future Work

MURAL-forests and the MURAL-embedding could potentially be used for supervised tasks such as classification, as indicated by the separation of low and high-risk patients. A potential framework for ensuring fidelity of the MURAL-embedding is the use of geometry regularized autoencoders and selecting embeddings with the lowest reconstruction error, and then using the embedding in a semi-supervised fashion as part of a feedforward neural network. This will be the focus of future efforts to use MURAL-embeddings for supervised tasks.

Conclusion

We present MURAL, a random-forest based framework for deriving distances between patients using mixed-model electronic health record data. We showed that the resultant MURAL-embeddings recapitulate the structure and heterogeneity of patient populations better than alternatives—thus paving the way for unsupervised learning to be used on clinical data. We note that most of the machine learning methods that are currently used for modeling clinical data require supervised training and large sets of annotated and labeled samples. However, by making clinical data amenable to unsupervised approaches, we can diminish this burden and even discover novel, clinical groupings of patients that could be meaningful for diagnosis, prognosis, or treatment.

Electronic Health Record Phenotyping Using Knowledge Graphs: Embedding Signals on Graph with Unbalanced Diffusion Earth Mover's Distance

Introduction

The task of comparing probability distributions is applicable to a wide variety of machine learning problems, giving rise to popular φ-divergences such as the Kullback-Leibler (KL), Hellinger, or total variation (TV) divergences, which ignore the underlying geometry of their support. The Earth Mover's Distance (EMD), also known as the Monge-Kantorovich or Wasserstein Distance, explicitly takes into account this underlying geometry via a domain-specific ground distance, which has many advantages on empirical probability distributions.[174,186] Here, we show that earth mover's distances are useful in a new domain: that of graph signals. In modern relational machine learning, we encounter large graphs that arise via interactions between entities in many domains.[187,188] Features of such entities can be considered as signals on the graph. For such signals, which often tend to be noisy, we propose a new unbalanced graph earth mover's distance, and use it to organize the signals and determine relationships between them.

Since graphs can contain tens (Cora)[187] to hundreds of thousands of nodes (SNOMED-CT)[189], there is a great need for this measure to be computationally efficient. While the Wasserstein distance is intuitively attractive, it presents computational challenges. Here, based on the recent diffusion EMD method[190], we show that an efficient unbalanced EMD between signals can be computed as the difference between graph convolutions of the signal with multiscale graph kernels. This unbalanced EMD can be computed in linear time with convergence guarantees and without solving an optimization problem. We call our distance unbalanced diffusion earth mover's distance (UDEMD).

While previous work on Wasserstein distance embedding mostly focused on its relation to the balanced optimal transport problem[176,191-194], we propose an unbalanced Wasserstein embedding approach between large number of distributions defined as signals on graphs. Since graph signals tend to be noisy, an unbalanced transport, which can choose not to transport parts of the data space when it is inefficient to do so, leads to more robust distances between graph distributions that are less sensitive to outliers in the signal.

We apply UDEMD to medical knowledge graphs using Systemized Nomenclature of Medicine - Clinical Terms (SNOMED-CT).[188] We show that unbalanced diffusion EMD can be used to find meaningful distances between patients which successfully clusters patients into different diagnosis categories, and allows us to find relationships between patient features. We also apply UDEMD to single cell RNA sequencing data where we can model both cells as signals on gene interaction graphs or genes as signals on cell-similarity graphs. In cases where the gene regulatory network is well known, researchers have shown that affinity between cells can be computed as an earth mover's distance.[195,196] We show that UDEMD runs orders of magnitude faster than the Sinkhorn and network simplex methods used in those works, while maintaining accuracy. In cases where the gene regulatory network is not well known, we model the transposed problem, deriving groupings of genes that function similarly by modeling genes as expression values over single cells. Here, we show that the UDEMD provides robust distances that recapitulate ground truth gene groupings in single cell data from peripheral blood mononuclear cells (PBMCs).

PRELIMINARIES

In this section we review the Wasserstein metric, embedding based methods for approximating it, and unbalanced optimal transport.

The Wasserstein metric is a notion of distance between two measures μ,v on a measurable space Ω endowed with a metric d(·,·) known as the ground distance. The primal formulation of the Wasserstein distance Wd, also known as the earth mover's distance, is defined as:

$$W_d(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} d(x, y)\pi(dx, dy), \qquad (1)$$

where Π(μ,v) is the set of joint probability distributions π on the space Ω × Ω, such that for any subset ω ⊂ Ω, π(ω × Ω) = μ(ω) and π(Ω × ω) = v(ω). Also of interest is the entropy regularized Wasserstein distance[175], which reduces the computation to O(n2). This algorithm is extremely parallelizable, and works quite well even for a small number of iterations[186], and there are many works investigating how to scale this to larger problems.

However, when comparing a large number of signals (say m), we must solve the optimization for each pair of signals, i.e. O(m2) optimizations. For this reason, we turn to methods that approximate the dual of the Wasserstein metric, also known as the Kantorovich-Rubenstein dual formulation, which relies on witness functions. Many works optimize the cost over a modified family of witness functions such as functions parameterized by neural networks[197-199], functions defined over trees[190,200], and wavelet bases[193,194]. An efficient algorithm recently proposed is Diffusion EMD[191], it is based on a multi-scale representation of the signals. Indeed, it can be seen as a weighted average of the L1 distances between two signals at different scales.

There are numerous formulations of unbalanced optimal transport both to accommodate problems with unequal masses and to provide robustness to outlier points.[174,201]. In general these can be formulated as a mixture between a pure optimal transport problem and a φ-divergence. We focus on the formulation using the total variation, referred to as the TV-unbalanced problem:

$$\text{TV-UW}_d(\mu, \nu) = \inf_s \left\{ \text{W}_d(\mu + s, \nu) + \lambda \text{TV}(\mu + s, \mu) \right\}, \quad (2)$$

where $\lambda = \min(\lambda\mu, \lambda\nu)$ and λμ,λν control the relative cost of mass creation / destruction compared to transportation. Intuitively, we can think of Eq. 2 as minimizing over the "teleporting" mass s, that is too costly to transport.

In the unbalanced optimal transport literature, most often considered is the KL-divergence formulation which can be solved efficiently in the case of entropic regularized problem[202-204], but is difficult to optimize stochastically as is possible in the balanced case, limiting scalability.[205,206] The TV-unbalanced problem (Eq. 2) can be solved by adding a "dummy point" that is connected to every point with equal cost.[207,208] However, adding a dummy point removes the metric structure necessary for dual-based Wasserstein distances. It is not immediately obvious that Eq. 2 is efficiently computable while maintaining this structure. To address this issue, Mukherjee et al showed that the TV-unbalanced problem can be solved through cost truncation.[209] Following their work, we will show that there is an embedding of distributions to vectors where the L1 distance between vectors is equivalent to the TV-UW between the distributions.

UNBALANCED DIFFUSION EARTH MOVER'S DISTANCE

While Diffusion EMD can provide an earth mover's distance between graph signals, its formulation is not motivated by considering noisy signals on graphs or outliers, but rather geared to avoid high dimensional density estimation.[191] Here, we focus on utilizing EMD to organize graph signals. Therefore, we are interested in distances that are immune to outlier spikes in the signals. While the multiscale smoothing proposed is effective in handling noisy perturbation of the signals, it is less effective at dealing with outlier vertex components of the signal. However, as we show here, the construction can be adapted to consider unbalanced transport, which is essentially based on the idea that a more faithful earth mover's distance is given by a transport in which we ignore some of the mass – particularly, mass that requires large transport costs. To incorporate this idea, we modify the formulation by only considering certain scales. This yields the Unbalanced Diffusion EMD (UDEMD), which is topologically equivalent to the total variation unbalanced Wasserstein distance.

Definition 1. The Unbalanced Diffusion Earth Mover's Distance (UDEMD) between two signals μ,v is

$$\mathrm{UDEMD}_{\alpha,K}(\mu,\nu) := \sum_{v\in V}\sum_{k=0}^{K} \|g_{\alpha,k}(\mu(v)) - g_{\alpha,k}(\nu(v))\|_1 \quad (3)$$

where 0 < α < 1/2 is a meta-parameter used to balance long- and short-range distances, and

$$g_{\alpha,k}(\mu(v)) := 2^{-(K-k-1)\alpha}\left(\boldsymbol{\mu}^{(k+1)} - \boldsymbol{\mu}^{(k)}\right) \qquad (4)$$

where μ(t) is short for μ and K is the maximum scale considered.

The scale K relates to the unbalancing threshold (see Fig. 1 and discussion in Sec. 3.1). In practice, α is set close to 1/2, hence we drop the subscript and use the notation UDEMDK.

3.1. Equivalence to (unbalanced) Wasserstein distance

In Pele et al, it was shown that truncated-cost optimal transport distances were equivalent to unbalanced Wasserstein distances, and that they are useful in outlier detection. [208] However, there the proposed implementation used a truncated matrix with the standard Sinkhorn algorithm.[175] Here we show a similar result for the Unbalanced Diffusion EMD from Def. 1, i.e., showing that with scale truncation it is equivalent to an unbalanced Wasserstein distance. We first adapt Theorem 3.1 from Mukherjee et al in the following Lemma 1, which will in turn be combined with Lemma 2 to yield this result. .[209]

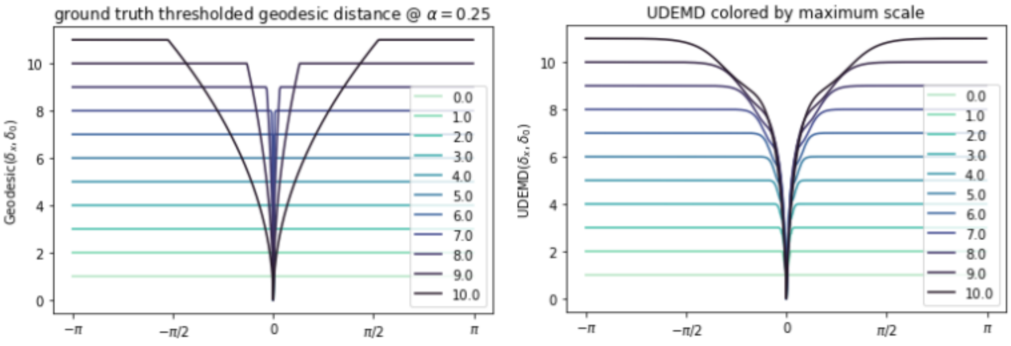
Lemma 1. *The Wasserstein distance with a truncated ground distance $d\lambda(x,y) = \min(\lambda,d(x,y))$ for some constant λ and distance d is equivalent to a total variation unbalanced Wasserstein distance for some constant λ, i.e., $Wd\lambda(\mu,v) = TV\text{-}UWd(\mu,v)$.*

The theory developed in Tong et al. assumed that the support of the considered distributions was a closed Riemannian manifold.[191] In such a case, Diffusion EMD will converge to a distance that is equivalent to the Wasserstein distance defined with the geodesic on the manifold.

The following Lemma extends this theory to show that UDEMD (Def. 1) will converge to a Wasserstein distance where the ground distance is a thresholded geodesic.

Lemma 2. $UDEMDK(\mu,v)$ approximates a metric equivalent to the Wasserstein distance $Wd\lambda(\mu,v)$, defined as in Lemma 1, with the ground distance being a truncated geodesic distance on the manifold, i.e., $d\lambda(x,y) = \min(\lambda,\rho(x,y))$ for $\lambda > 0$.

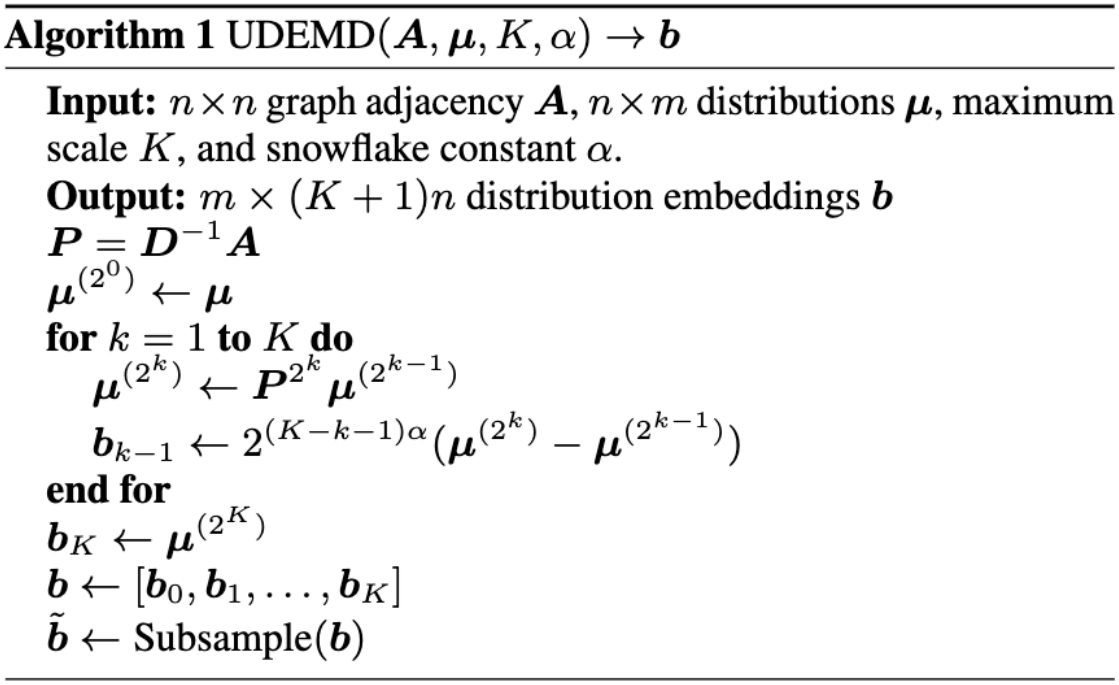
Proof. We present a proof sketch here; the main part of the proof follows the same lines as in Corollary 3.1.[191] In Def. 1, an anisotropic kernel P is used, which can be shown to converge to the heat kernel on a Riemannian manifold (Prop. 3).[210] In Leeb and Coifman, it is shown that the construction of Def. 1 using the Heat Kernel will converge to a metric that is equivalent to the Wasserstein with ground distance $\min(1,\rho(x,y)2\alpha)$, where ρ is the geodesic on the manifold.[211] Because the metrics $\min(1,\rho(x,y)2\alpha)$ and $\min(\lambda,\rho(x,y)2\alpha)$ are equivalent for $\lambda > 0$, the Wasserstein distances induced by these metrics are also equivalent.



(a) Thresholded geodesic (b) UDEMD between diracs

Figure 26: On a ring graph n = 500 compares the UDEMD to the thresholded ground distance, this suggests that UDEMD closely approximates the thresholded ground distance with λ ≈ 2K.

By combining Lemmas 1 and 2, we have that the UDEMD from Def. 1 approximates a metric equivalent to an unbalanced optimal transport metric. Formally, using the equivalence notation from, we have UDEMDK(µ,v) ' TV-UWd(µ,v). We note that while our result here establishes a relation between these two metrics, it does not directly quantify the relation between the λ and K. We leave careful theoretical and rigorous study of this relation to future work, but mention here that we observe empirically, as shown in Fig. 1, the choice of K indeed acts in a similar way to the threshold λ on the ground distance.



```
Algorithm 1 UDEMD($\boldsymbol{A}, \boldsymbol{\mu}, K, \alpha$) $\rightarrow \boldsymbol{b}$
  Input: $n \times n$ graph adjacency $\boldsymbol{A}$, $n \times m$ distributions $\boldsymbol{\mu}$, maximum scale $K$, and snowflake constant $\alpha$.
  Output: $m \times (K+1)n$ distribution embeddings $\boldsymbol{b}$
  $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$
  $\boldsymbol{\mu}^{(2^0)} \leftarrow \boldsymbol{\mu}$
  for $k = 1$ to $K$ do
      $\boldsymbol{\mu}^{(2^k)} \leftarrow \boldsymbol{P}^{2^k}\boldsymbol{\mu}^{(2^{k-1})}$
      $\boldsymbol{b}_{k-1} \leftarrow 2^{(K-k-1)\alpha}(\boldsymbol{\mu}^{(2^k)} - \boldsymbol{\mu}^{(2^{k-1})})$
  end for
  $\boldsymbol{b}_K \leftarrow \boldsymbol{\mu}^{(2^K)}$
  $\boldsymbol{b} \leftarrow [\boldsymbol{b}_0, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K]$
  $\tilde{\boldsymbol{b}} \leftarrow$ Subsample($\boldsymbol{b}$)
```

To compute the UDEMD defined in Def. 1, we present Alg. 1 with time complexity O(2Km|E|), which is similar to algorithms used in graph neural networks. Our algorithm scales well with the size of the graph, the number of distributions m and number of points n, but poorly with the maximum scale K. We note that the maximum scale considered for Diffusion EMD was of order O(log|V |), derived from the convergence of the heat kernel to its steady state. Here, on the other hand, we decouple the tuning of K and find that a much smaller maximum scale suffices, and in fact (as discussed in Sec. 3.1) corresponds to a well characterized unbalanced earth mover's distance on the underlying geometry of the graph. This leads to Alg. 1 emphasizing preferable scaling properties for small K, and easily accelerated by computation on GPUs.

Results

In this section, we show that UDEMD is an efficient and robust method for measuring distances between graph signals and then using the distances to find embeddings and organization of the signals (often entities such as patients). We compare UDEMD to a GPU implementation of numerically stabilized Sinkhorn optimization that includes minibatching of sets of distributions. However, despite this, this method runs out of memory when there are beyond 10,000 nodes in the graph. We note that all methods of this type require solving m2 optimizations, even when looking for nearest neighbors. Unless otherwise noted, we set K = 4 and α = 1/2.

Spherical data test case

To test the speed and robustness of UDEMD we begin with a dataset where we have knowledge of the intrinsic ground distances and can vary the number of points and distributions. For this dataset we sample m Gaussian distributions with means distributed uniformly on the unit sphere with 10 points each for a total of n = 10m points. We add a random noise spike at a uniformly random location on the sphere to check robustness to this type of noise. The goal is to predict the neighboring distributions on the sphere. We find that UDEMD is significantly more scalable and find that there is a sweet spot in terms of K at K = 4 for this dataset. The UDEMD with K = 4 performs significantly better than the balanced Diffusion EMD case with this type of noise. This supports the claim that setting is beneficial in real world datasets. UDEMD also outperforms the graph-TV distance as it is both faster and more accurate at K = 1, and more accurate overall. (Figure 26)

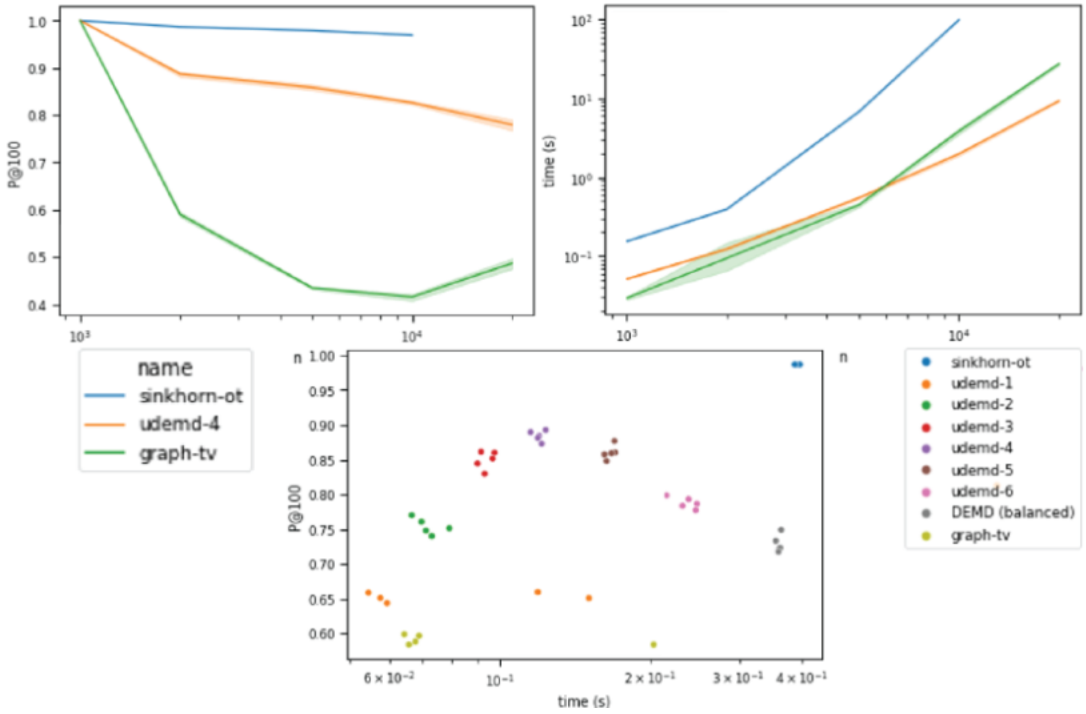
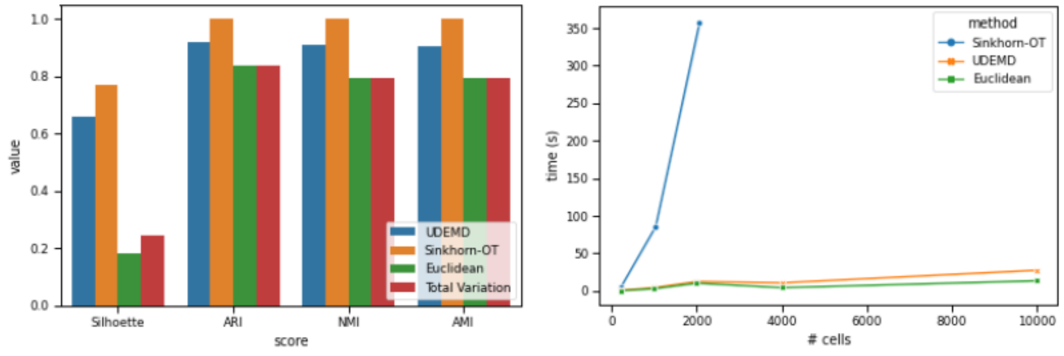


Figure 27: UDEMD is more scalable than Sinkhorn-OT and performs better than graph total variation. (left) Shows performance as measured by P@100, the fraction of the 100 nearest

neighbors predicted correctly, against problem size. (middle) Shows time against problem size, and (right) shows performance vs. time on a problem size of n = 2000 for different choices of K.

Single-cell data with cells as signals over gene graphs

We consider 206 cells from the K562 human lymphoblast cell line as signals over a known 10000-node gene graph in single-cell RNA seq data [27]. We measure the distance between cells based on their transport on this gene graph. This was recently independently proposed by [10] and [11], who showed that OT over the gene graph can provide better distances between cells than Euclidean measures. We measure the performance of these methods based on how well the resulting distance matrix between cells matches the clusters according to four scores: Silhouette score, the adjusted rand index (ARI), the normalized mutual information (NMI), and the adjusted mutual information (AMI). In Fig. 3, we see that UDEMD performs almost as well as Sinkhorn-OT, and much better than the Euclidean and total variation distances that do not take into account the gene graph as well as much faster than Sinkhorn-OT, scaling almost as well as Euclidean distances due to the embedding. Note however, that using balanced transport (see Fig. 2 right) degrades the accuracy. The balanced transport compared here is the original Diffusion EMD from [5] which is not a thresholded distance, and thus noise in the data are able to perturb the accuracy of the distances. (Figure 27)



(a) Performance metrics (b) Computation comparison

Figure 28: UDEMD achieves better clustering than Euclidean and total variation (TV) distances, and performs similarly well to Sinkhorn-OT but is much more scalable with similar scalability to

Euclidean and TV distances. (a) performance in terms of Silhouette score, ARI, NMI, and AMI (b) computation time vs. problem size.

Single-cell data with genes as signals over cell graphs

Next we applied our approach to 4,360 peripheral blood mononuclear cells measured via single cell RNAseq publicly available on the 10X platform. We consider three curated gene sets that are explanatory for this dataset. We compare the distances between genes using UDEMD, Euclidean, total variation and Sinkhorn-OT distances. We can see that the genes canonical for monocytes (orange), T cells (green) and B cells (blue) all appear to be closely positioned to one another and separate between the groups in our embedding in contrast to a Euclidean distance embedding of the genes where the clusters are less clear (Fig. 4a). Visualizing the UDEMD distance between our 46 genes in a heatmap, we can identify the three clusters as dark blocks of low distance (Fig. 4b). This result is quantified in (Fig. 4c), where UDEMD performs the best on 3/4 metrics. Last, we tried to see how diffusion time scale impacted silhouette score, identifying that score maximized at a timescale of K = 10 and did not improve with higher scales. (Figure 28)

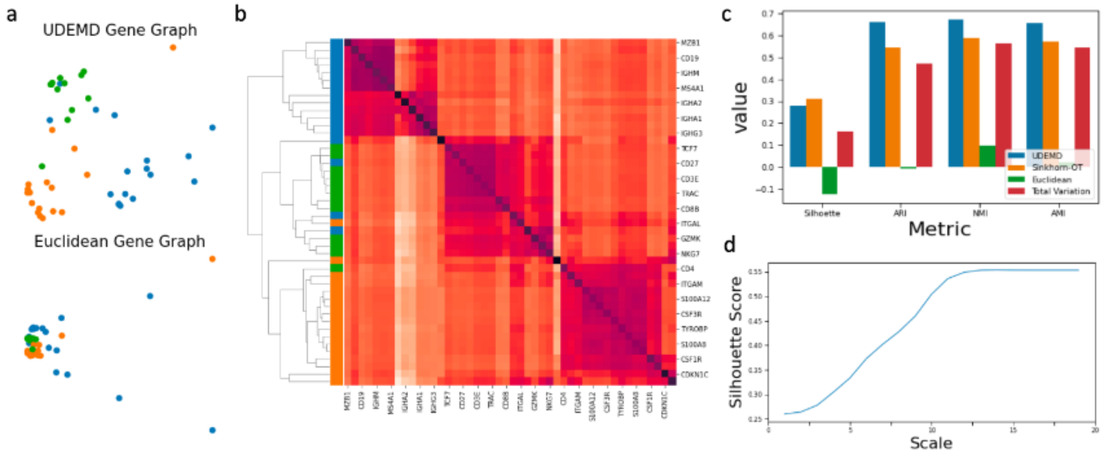


Figure 29: (a) Visualization of gene graphs of 46 genes canonical for different cell types using UDEMD and Euclidean ground distances (blue for B cells, orange for monocytes and green for T cells), (b) heat map of gene distances (c) clustering performance (d) silhouette score vs. maximum diffusion scale K.

A Patient Concept Knowledge Graph

We consider a knowledge graph constructed from medical concepts captured in clinical documentation and reporting. SNOMED-CT is a widely used collection of terms and concepts with defined relationships considered to be an international standard for medical concepts captured from the electronic health record. SNOMED-CT has a pre-defined knowledge graph with concept-relation-concept triplets, which we subset to the Clinical Findings concept model (version 3/2021). We used 52,150 discharge summaries from MIMIC-III, which contain all information about a patient's hospital course and extracted concepts using MetaMap (version 2018) [28]. These medical concepts were then used as signals on the SNOMED-CT knowledge graph, which link all relevant concepts together. The metadata used to label patients included primary diagnosis, a physician-designated diagnosis which was stored separately in the MIMIC-III database.

One of the advantages of the UDEMD-based embedding is the identification of clinically meaningful overlaps that may not be apparent from the single primary diagnosis recorded in the database. Patients with a primary diagnosis of intracranial bleeding (bleeding in the brain) can also have primary brain masses and tumors. Compared to the spurious fragmentation of patients with the same diagnosis of intracranial bleeding into several clusters in the TV embedding, UDEMD consolidated patients with the same diagnosis of intracranial bleeding and specifically grouped those that may have had bleeding due to a primary brain mass or tumor (See Figure 29B). Interestingly, UDEMD also identified patients who were predicted to have intracranial bleeding to have the diagnosis of stroke with higher accuracy, reflecting consistency with the fact that a subtype of stroke (hemorrhagic) is due to intracranial bleeding (Figure 29D).

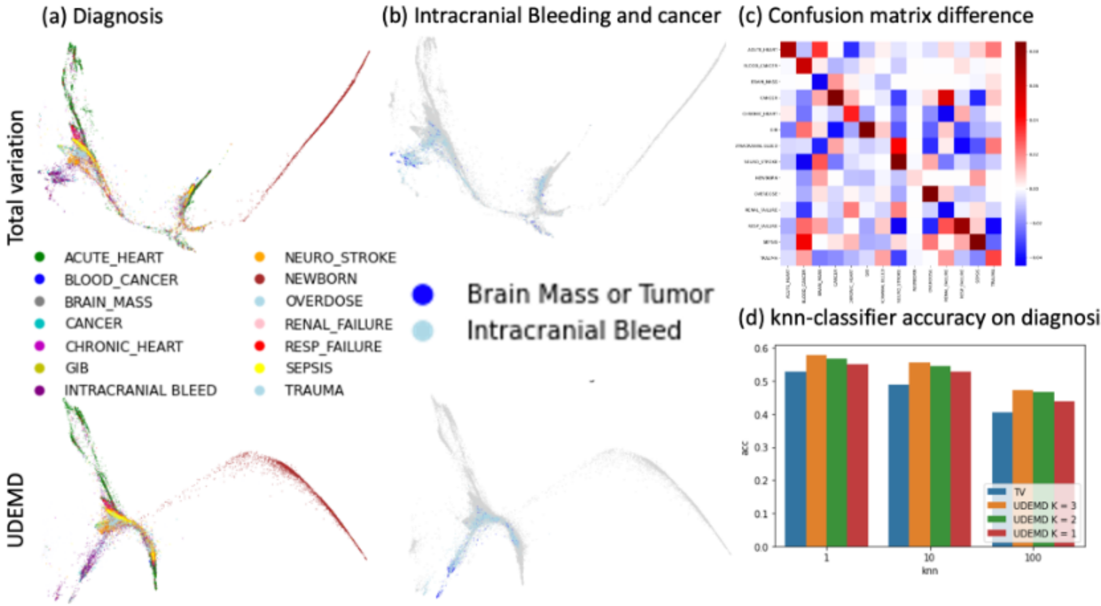


Figure 30: Embeddings of patients modeled as signals over the SNOMED-CT graph using TV distance (a top) and using UDEMD distance (a bottom), colored by patient diagnosis. UDEMD better organizes the space as noted by selected terms in (b), difference of confusion matrices in (c) and k-nearest neighbors classification accuracy on the diagnosis in (d). In (b) note that the TV embedding (top) creates a spurious separation (due to noise in the signal) between subsets of patients who display intracranial bleeding that is not distinguished by diagnosis. On the other hand, the UDEMD embedding (bottom) shows a continuum of patients with this diagnosis. The same holds for patients with brain mass or tumor shown in green.

Conclusion

In this work we explored the use of earth mover's distance to organize signals on large graphs. We presented an unbalanced extension of Diffusion EMD, which we showed approximates a distance equivalent to the total variation unbalanced Wasserstein distance between signals on a graph. We showed how to compute nearest neighbors in this space in time log-linear in the number of nodes in the graph and the number of signals. Finally, we demonstrated how this can be applied to entities which can be modeled as signals on graphs between genes, cells, and biomedical concepts.

Data Access, Bias, and Equity in Machine Learning Interventions

Introduction

All data are not created equal. The ideal for data is to reflect reality, or what is actually happening. Algorithms are well-defined procedures for carrying out computational tasks. Algorithms are the main vehicle for machine learning interventions, which may aim to provide a diagnosis or give a prediction to guide clinical decision-making. The underlying data used to train and validate these tools must be carefully considered. Data access, data bias, and data equity are three areas that directly impact the use of and trust in machine learning interventions. Data access refers to the unrestricted ability of a designated party to view, test, and manipulate the raw data used to train machine learning algorithms and can be a proxy indicator for the responsible party when algorithms err. Data bias identifies areas in the data that may not reflect best ethical or clinical practice, and thus settings or populations where machine learning interventions should be interpreted with caution. Data equity involves the concept of algorithmic fairness, which could ultimately impact provider trust in machine learning interventions. For providers using machine learning algorithms, data access provides a reference for who should be accountable for poor performance, while data bias and data equity help providers interpret and trust algorithmic output when applied across different settings in clinical care.

The majority of data used for clinical and translational machine learning tools can be captured in the electronic health record as part of routine clinical care. However, additional sources of data such as patient-generated data through wearables, mobile applications, and ambient sensors are increasingly prevalent in the healthcare data ecosystem. While the specific relevance of access, bias, and equity may differ across these different data modalities, there are also shared general principles when using health-related data.

Data Access

Healthcare data is sensitive and subject to special restrictions to ensure security and privacy. On a high level, the governance structure for “big data” or “artificial intelligence” is important in understanding the parameters of data access across societies. For clinical providers using machine learning interventions, data access refers to the entity with the permission to examine the underlying raw data used to train the algorithm. As such, if a decrease in performance is noted with errors, the entity with access to the underlying data is thus responsible for identifying the root cause of the error.

While governance encompasses the entire life cycle of machine learning-based tools, the way that governance is structured directly impacts data access. Industry self-governance differs from government conceptions, which may vary by level (national government, supragovernmental organizations) and region.[212-214] While there is limited information about the consensus to data access across the different settings, a representative publication about industry self-governance does not mention data access specifically but assents to the best practice of data transparency and reporting. For U.S. governmental FDA guidance, no explicit reference is given to data access but rather voluntary collaboration when piloting machine learning-based interventions to assess real-world performance. A paper written by Chinese authors proposed a framework for big data governance for health information networks does not explicitly address data access, but in a series of guidelines suggests centralization and integration on a national scale in conjunction with the healthcare industry. In contrast, the European Union Proposal for Regulation Laying Down Harmonized Rules on Artificial Intelligence clearly articulates the data access for training, validation, and testing dataset to be given to market surveillance authorities.

For clinical providers in different regions of the world, it may be instructive to consider that issues with performance requiring access to the underlying data may lay with industry firms in the United States, the market regulatory authority in Europe, and the central government agency in China.

Security and Privacy

Data access is balanced by the need for security and privacy, particularly for healthcare data. The Health Insurance Portability and Accountability Act (HIPAA) applies specific requirements to all projects that are considered human subjects research, and the EU Data Privacy Regulation

(GDPR) is a broader set of regulatory guidance that has specific protections for personal data. HIPAA-related data protections include informed consent (or an explicit waiver with justification) and deidentification with safe harbor (removal of 18 specified personal identifiers) and expert determination. In light of these requirements and concerns about the risk of re-identification, measures to maintain data security and personal privacy are an essential part of granting data access. One framework has been proposed by the National Institute of Standards and Technology (NIST) at the U.S. Department of Commerce to manage the risk of healthcare data: Identify, Protect, Detect, Respond, and Recover. To ensure that only persons or entities with adequate training and credentials have access to sensitive data, it is important to identify data, personnel, devices, systems, and facilities pertaining to health-related data. By cataloging, maintaining an inventory and mapping data flows as well as the roles and responsibilities for the workforce healthcare organizations can steward access and manage the risk of a data breach or unauthorized access. Equally important is the role of protecting access using identity management, authentication and access control enabled by security protocols and contingency planning. When a breach of privacy occurs, whether through inadequate deidentification or adversarial attack, detecting the incident through continuous monitoring and planned, coordinated responses with mitigation strategies are key in containing the damage. Finally, recovery of data and planning to incorporate lessons learned into the existing infrastructure can help safeguard the system from future incidents. For special populations, such as veterans receiving care through the Veterans Affairs Department of the United States, the challenge of using sensitive data to drive innovations that may improve care delivery may have additional barriers to data access.

Data Access for Research Endeavors

While data access is generally relevant to practitioners as users of machine learning interventions, researchers also seek data access for algorithmic development or independent validation. While access to institution-specific datasets or publicly available deidentified datasets is possible, researchers may seek more representative datasets that are not limited to a specific center of health system. Two datasets that have well-defined security protocols and rich longitudinal data include the All of Us Research Program, an initiative by the National Institutes of Health, and the UK Biobank are high profile efforts to build a diverse health database representative of modern U.S. and U.K. society, respectively. When thinking through data access, the mission of All of Us encompasses not only university researchers, but also citizen scientists and administrators interested in quality improvement initiatives. The UK Biobank furthers the access to specifically include those in low-middle income countries who do not traditionally have the capability of acquiring such large-scale and in-depth data.

Data bias

As the relationship between patients and physicians persists beyond the initial meeting and diagnoses, the data framework should go from thinking about data just for initial model development and validation to encompass the product lifecycle.[215] A key question that should be asked throughout the lifecycle is the following: do choices made about the data that is measured or captured during the lifecycle worsen or perpetuate existing health inequalities?[216] As the Oslerian ideal of equanimity, or mental equilibrium, motivates providers to consider clinical findings above and beyond the temptation to fall into specific cognitive biases, this question should motivate physicians, researchers, and data analysts to go from focusing on getting as much data as possible to critically considering the characteristics and deficiencies of the data being used.

Electronic Health Record Data

Data collected in the electronic health record or as part of routine clinical care is generated in the context of clinical medicine, where the patient-physician relationship should be considered as a framework for understanding bias. In this framework, the patient may not tell the physician all the relevant information, may seek care with other physicians, and may not have the means to undergo the recommended testing. The physician may have a unique diagnostic or treatment style and may perceive patient reports through the lens of their experience and perception.

Finally, the historical practice of medicine has incorporated race explicitly in ways that purposely bias calculations of clinically significant measures.

Missing data

Missing data can be a source of bias, but is only addressed or accounted for in 54% of predictive algorithmic studies using electronic health record data.[24] Missingness could be due to absence of provider entry into the health record, an error of omission, or secondary to lack of access to diagnostic tests or procedures.[216,217] Lack of access has been documented in surgical care and endoscopic access, particularly impacting patients identifying as Black.[70,218] Fragmentation of care without systemic interoperability can contribute to data missingness, particularly if they are linked to other societal factors such as low socioeconomical status, psychosocial issues, or immigration status.[219,220] In particular, demographic and socioeconomic data are often incomplete, with an estimated one-third of commercial insurance plans reporting complete or partially complete data on race.[220]

Variation in Practice Patterns

Since practice may change across different providers and health systems, differential care patterns can lead to misclassification and measurement error.[216] This may be seen across teaching and non-teaching settings, or across urban and rural areas and correspond with uninsured patients or patients on Medicaid, which has been well documented in the Emergency Department setting.[221]

Provider Bias When Interpreting Patient Reported Symptoms

Patient language is usually communicated and filtered through providers who then enter the information into the electronic health record. As such, the physician may selectively record information according to their clinical experience, and unfortunately sometimes their bias. In Psychiatry, where patient-reported data is exclusively filtered through providers, this has been seen as potentially problematic in the stage between the expression of data by the patient and the interpretation of data by the provider; data captured by the provider and then analyzed by a machine learning algorithm to predict an outcome can magnify the bias.[222] For example, if providers prescribe higher doses of psychiatric medications to patients with minority race backgrounds despite having similar reported symptoms, an algorithm could then be biased to predict higher doses of necessary medication by race. Another example can be that since women are more likely to receive personality disorder diagnoses compared to men when presenting with the same symptoms of trauma, an algorithm may perpetuate this diagnostic pattern as predicted diagnoses.[223-225]

Explicit Racial Correction in Clinical Tools

Corrections in definitions of organ performance, such as the glomerular filtration rate, heart failure mortality risk, and pulmonary function tests, may be based on race. For example, Vyas et al recently identified 13 clinical algorithms in which race was explicitly used to modulate risk and determine courses of treatment.[226]

Other Data Sources: Patient Wearables and Ambient Sensors

Data captured from electronic health records is primarily filtered through provider entry and health system priorities, whereas wearable and app-captured data reflect direct patient input and ambient sensors' monitoring of physical environments. However, the issue of bias can also impact these data sources along similar lines of resource access due to socioeconomic status and optimization of these devices for specific populations. In digital behavior change interventions to increase physical activity, the devices appear to have greater effectiveness for people with higher socioeconomic status.[227] For wearables tracking health-related information (e.g., heart rhythms, sleep patterns), the challenge of the sensor technology for maintaining accuracy across skin pigmentation tones as well as absence of diverse representation in validation studies may lead to differential effectiveness by race and ethnicity.[228]

For ambient sensors in health settings, the measurement or capture of data on multiple participants in different contexts can have the potential for reinforcing existing biases. Behavior between participants in a certain context may have a specific interpretation that may bias the interpretation of algorithms trained on this setting when transferred to another setting.[229] For example, an sensor-based algorithm trained to identify provider activity in an intensive care unit may not be directly applicable to a psychiatric ward or maternity care unit.

Data equity

Fairness for algorithms has been defined by three pillars: transparency, impartiality, and inclusion.[230] Practical categories such as geographic region, socioeconomic strata, gender, and race/ethnicity can be used as a starting point for thinking about overall fairness. Explicitly addressing these categories may provide a basis for provider trust in the applicability of these tools in practice.

Transparency includes interpretability, explainability, and accountability. While interpretability can be included in basic standards set by regulatory bodies, interpretability for providers may include specific aspects of which data elements are used for the algorithmic output can be relevant for providers using machine learning interventions. In particular, explainability can put those data elements in the specific context from which the data is generated, which can give providers additional information about the relevance of the algorithmic output. By understanding the setting from which training data is drawn, then the factors used to generate the prediction can be taken into account by providers using the tool. For example, a machine learning intervention to predict outcomes for patients with myocardial infarctions may be influenced by the presence of disparity in survival for women compared to men when presenting with myocardial infarctions; additionally, another intervention to predict perinatal mortality may amplify the bias of mortality for Black women due to the underlying data trend increased perinatal mortality for Black women in the US.[231-233] Finally, accountability is crucial, since the question of responsibility in the event of an adverse outcome must be defined to mitigate the risk to providers using these tools. When thinking through accountability in healthcare, it may be useful to think of the machine learning intervention in terms of a consultation to access additional information or expertise they cannot otherwise access. In this context, a particular challenge is navigating the presence of deep expertise held by providers, who may have specific concerns or questions. There should be a clear designation of responsibility in the event of an error, so that providers know who to consult to ensure that the error is not repeated.

Impartiality includes provenance (the origins and characteristics of the data) and implementation. The starting point is clinical relevance in the specific setting of use, to ensure that the patient population in the intended setting is represented by the data used to train the machine learning intervention.[234] Then, during implementation the anticipated harms should be considered to evaluate if there is disproportionate impact on specific populations. This process should include evaluation of discriminatory practices that arise or are exacerbated by the integration of the machine learning intervention. For example, in a modeling analysis machine learning based predictive tools for medical appointment scheduling may amplify the higher no-show probability for black versus non-black patients and lead to wait times up to 30% longer than non-black patients due to recommendations for overbooking.[235]

Inclusion encompasses data completeness and utilization of traditionally excluded data sources, such as patient-reported or community-reported data. Data completeness should consider geographic distribution and representativeness, which is currently skewed in machine learning applications for clinical medicine to disproportionately use cohorts from California, Massachusetts, and New York.[236] Race-based differences may be more relevant in societies with historical policies of purposely disadvantaging specific racial groups, though other categories may better reflect the specific historical disparities of each society.

Algorithmic Stewardship as a Framework for Mitigation Strategy Against Bias and Inequity

The essential part of strategies to identify and address issues with algorithmic bias is the human in the loop. While data access is essential for any mitigation strategy, the study and monitoring of real-world effectiveness of algorithmic machine learning interventions is arguably more important in identifying areas of adverse effect possibly attributed to bias or inequity. Algorithmic stewardship is a useful framework to encompass necessary practices to mitigate bias and test for inequity.[237] Algorithmovigilance, which refers to methods for evaluation, monitoring, understanding, and preventing adverse effects of healthcare algorithms, is a useful framework for the development and deployment of informatics-based methods for debiasing algorithms.[238] The main parts of stewardship include creation and maintenance of algorithm inventories, an auditing process prior to deployment, and periodic review by a group with oversight responsibility. While this can be thought of at a health-system level, it is likely that governmental or supragovernmental support may be necessary due to the limited expertise available at a health system level to understand and evaluate algorithmic performance with clinical relevance. The need for an interdisciplinary approach, including sociological and ethics expertise, is critical, since algorithms are typically designed to maximize performance and efficiency and not to reflect human values. Imposing human values can be directly at the cost of efficiency and be applied judiciously when the benefit outweighs the cost.[216]

Future Challenges for Development and Application of Machine Learning Models to Clinical Care

Introduction

Clinical management of gastrointestinal diseases span the spectrum from acute to ongoing chronic care and use multiple types of information including endoscopic video, radiologic imaging, manometric readings, and genomic data. With recent advances in artificial intelligence to process imaging, text and genomic data, there is great promise for AI-assisted tools to advance the care of patients with gastrointestinal diseases. However, given the complexity of clinical care, there are significant logistic, regulatory, and ethical challenges in determining appropriate and optimal use of the technology.

The Potential of Artificial Intelligence and Machine Learning in Analyzing "Big Data"

Artificial intelligence is a field that has advanced rapidly in the age of increased computational power, algorithmic sophistication, and availability of data. There is a distinction between general artificial intelligence and narrow artificial intelligence. General artificial intelligence is theoretically identical to human intelligence and is not restricted to specific tasks. General artificial intelligence currently does not exist, but there are prototypes in natural language processing that appear to be a promising step in that direction (a new language generator called GPT-3 recently released by OpenAI in June 2020).

Machine learning, a subset of artificial intelligence, is a set of computational tools used in narrow AI applications, where the algorithms are trained to perform well for very specific tasks (e.g. identifying polyps on screening and surveillance colonoscopies). The advantage of machine learning over conventional statistical tools is the ability to analyze "big data", defined as datasets that are large (volume), complex (variety), and constantly updating (velocity). For medicine, the explosion of available data has been estimated as a doubling every 73 days, and machine learning tools are suited for analyzing the data to be used for diagnostic and prognostic purposes.[239]

Challenges facing AI implementation in Medicine

Machine learning is, first and foremost, a tool to be used in clinical care. Like any tool, the purpose must be judiciously and thoughtfully considered prior its use. Machine learning tools depend heavily on the data used in training and development, and may include mathematical and statistical assumptions that are unfamiliar to most clinicians. Logistic challenges can be categorized into understanding the care delivery process, data management, and algorithmic understanding. However, the overall environment for integrating AI-assisted tools needs more development to promote wider uptake. This includes regulatory guidance, standardized payment, and ethical challenges in data privacy, equity, and fairness.

Challenge 1: Understanding the Care Delivery Process

Before any machine learning tool is considered, a deep understanding of the problem and associated care delivery process is the key to any application of artificial intelligence to clinical care. The starting point to apply machine learning to clinical care should follow the framework suggested by Isaac Kohane: is the task simple or complex?[240,241] The task includes the clinical question, but also defines specific areas of the clinical process that can be optimized. By defining specific areas of the clinical process that can be optimized using artificial intelligence as a tool, the maximal benefit and value can be achieved for patient care and provider satisfaction. On a practical level, depending on the specific clinical problem and care process, the type of algorithm can be selected according to the required level of performance and amount of data available.

There is a growing recognition of the critical role of implementation into the clinical process for artificial intelligence tools, with the goal to “design the best possible care delivery system for a given problem.” [242,243,244] This is usually an iterative process that goes through the delivery process before, during, and after implementation with the AI tool and focuses on designing and improving user interfaces.

Challenge 2: Data Management

Data management is critical for artificial intelligence, because as the name machine learning suggests, models must have robust data to successfully learn the relevant patterns.

The first challenge that must be addressed is the availability of high quality data that is readily captured and accessible and can be generated with each iteration.[240] The principle “Garbage in, Garbage out” captures the core concept of ensuring that high quality data is used to train and test algorithmic performance. While advances in algorithmic development may help, the basis of most algorithms rest on the data itself.[132] The implications for data management does not end after training and testing; once the algorithms are trained and tested on high quality data, there should be a pipeline of consistently labeled data that can be used to continuously train the model in the “virtuous cycle” of the data ecosystem. To address this, many healthcare systems have worked towards data standardization and interoperability across platforms to ensure that imaging and clinical data can be pooled and used to generate consistent results.

Another challenge is data bias, or errors, that can lead to predictions that worsen incorrect practice patterns and may unintentionally contribute to worsening disparities in clinical care. For example, if an algorithm to predict risk of hospital-based intervention for patients presenting with overt gastrointestinal bleeding is trained in a setting where there is the wrong practice of overtransfusion, the algorithm may recommend admission for patients who could be discharged for outpatient care. Also, if a clinical dataset have gaps in data from vulnerable or underserved populations, unfair attitudes and practices may contribute to disparities in the output of the algorithm.[216] Although the potential for bias will always exist, rigorous validation can mitigate its effect on algorithmic performance. Ideally, the study designs for validation should include internal validation, external validation, calibration, and appropriate statistical testing that compares model performance with a control. External validation in particular is critical for both clinical data and imaging data, since the bias can be mitigated when data is pooled from multiple patient populations, centers, contexts, and manufacturers.

Prospective studies of deployed AI tools with iterative feedback and monitoring can identify areas of bias that can then be corrected. Clinical trial designs for artificial intelligence tools have been proposed based on the type of task, including randomized controlled trials, random tandem trials, A/B testing, and QI experimental designs.[245] Currently, new guidelines have been proposed for clinical trial protocols (SPIRIT-AI) and randomized controlled trials (CONSORT-AI) for AI interventions that can be used to design rigorous, high-quality studies.[245,246]

Finally, ongoing data maintenance is important to ensure that changes in patient populations or clinical care do not affect performance of the AI-assisted tool. Monitoring performance and re-training should be considered from the beginning, since there may be difference in data trends reflecting differences in patient outcomes due to dataset shift (practices, populations evolve over time), new therapies, or evolving epidemiology. Algorithmic stewardship is a new concept on a systems level that includes maintaining an inventory of existing algorithms, having regular audits of safety and fairness, and constant performance monitoring to prevent degradation over time.[154,247] As new data updates the AI, it will learn according to the input data and outcomes observed. As it optimizes to learn the pattern of data that predicts the outcome, it is possible that the performance will either improve in learning the actual predictors or deteriorate due to biases in the data leading to misclassification. Ongoing expert surveillance is key in troubleshooting the issue as either due to poor-quality input data or algorithmic error. With re-training and careful algorithmic monitoring the model should have equal if not better performance over time. The burden of maintenance and monitoring should ideally not be on the medical institutions or users, but rather the third-party vendor who should take responsibility for setting up the datastreams for regular updates, perform checks for algorithmic maintenance, and have protocols to investigate when an algorithm misfires. Furthermore, cost effectiveness should be assessed at both the institution level and across national and international boundaries. For AI-enhanced polyp leave-in strategy compared to resect-all-polyps strategy, one cost effectiveness study estimated savings of 18.9% and US$149.2 million in Japan, 6.9% and US$12.3 million in England, 7.6% and US$1.1 million in Norway, and 10.9% and US$85.2 million in the United States.[248] However, this study does not include a comprehensive evaluation of the cost of implementation of process changes including increased time for each procedure and actual real-world pattern of endoscopists using the tool.

Challenge 3: Algorithmic Understanding

Algorithmic interpretability is particularly important in clinical care, where providers have developed a deep expertise that can take into account factors that may not be captured by the machine learning model.[154] In gastroenterology, practitioners are specialists who are experts in the field should have the ability to verify the system performance. Use of the AI tools must consider the balance of power, in particular how the AI tools may impinge on professional authority for clinicians.[243] Furthermore, by understanding how the prediction is made, practitioners would be able to assess if the prediction is being generated from actual signal or is being distorted by confounding variables. Finally, the generated patterns can be tested and integrated into current scientific understanding to advance clinical care more generally.

Challenge 4: Algorithmic Adversarial Attacks

Adversarial attacks that exploit weaknesses in current algorithms by manipulating the input data is an emerging challenge that has particular importance for healthcare.[249] Currently the most active area in the United States is regarding insurance claims approvals via billing codes, since insurance companies deploy machine learning models for classifying certain claims. In particular interest for gastroenterologists, however, is the potential for visually imperceptible “adversarial noise” added to images that can cause deep learning models deployed on imaging to misdiagnose or miss pathology. Proposed measures to defend against attacks include backups that provide a “fingerprint” of data to be extracted and stored immediately after capture. This can then be used to compare to the image used for analysis to evaluate for data tampering and also be used to build resiliency in algorithms during real time deployment.

Challenge 5: Regulatory Guidance

Regulatory guidance is underdeveloped globally. Despite strong efforts on a national, regional, and international level to develop frameworks to ensure quality control and patient safety, there is still a high amount of uncertainty regarding the requirements that will be enacted.[250] Quality control through regulation is being developed through the International Medical Device Regulators Forum (IMDRF), the United States has proposed a new regulatory framework of Software as Medical

Device (SaMD) through the Food and Drug Administration, the European Union has proposed General Data Protection and Regulation (GDPR), and the China State Council has proposed a development plan for artificial intelligence.[250]

Challenge 6: Liability and Legal responsibility

Due to the regulatory ambiguity above, liability and reimbursement is not clearly defined. The issue of liability is critical for both firms developing AI-assisted tools and for medical provider end-users. If patients experience an adverse event based on clinical decisions made with AI-assisted tools, who is accountable? As a "black box" tool, the liability of adverse events to patients based on decisions made using AI-assisted tools may be shouldered by the manufacturer. If the output is sufficiently interpretable, however, the liability would likely be borne by the medical provider who made the clinical decision. Reimbursement should ideally compensate whichever party shoulders the risk; however, AI-specific reimbursement is notably absent from national healthcare systems and private payors. Currently, there is only one instance of payment specifically for use of an AI tool through the Centers for Medicare and Medicaid under the framework of new technology add-on payments, upcoming for fiscal year 2022. No other system-wide payment mechanism currently exists for reimbursement for AI tools, particularly for endoscopic enhancement.

Challenge 7: Ethical issues

Ethical challenges concern the interaction of these algorithms with human health and the safeguards that should be put in place to mitigate the potential adverse effects of these algorithms. Challenges include defining the role of informed consent in data utilization, maintaining privacy compliance across the spectrum of data users, returning results from analyses using patient data, and addressing equity in algorithmic development. Informed consent is a cornerstone of medical research, which recognizes the autonomy of patients and their right over their medical data. However, a challenge that should be considered in AI is the potential for using the same patient data for both specific conditions (e.g. inflammatory bowel disease), and also in aggregate (e.g. if the data is sent for epidemiological purposes).[251] Privacy is also challenging to maintain when AI may include a host of third party partners, including vendors, software developers, data scientists, and other systems. In particular, the United States requires HIPAA compliance across the spectrum of data users, and thus it is important to consider how to deidentify data and maintain a secure dataflow.[251] One specific area where this is important is when considering how and to whom results of the AI tool should be returned. This has implications for shared decision-making for patients and providers, since the findings may impact how the patient thinks about the next step in their treatment plan. By considering who should have access, what should be shared, and which threshold should be used to share specific findings, the patient and provider can hopefully use the AI tool to assist in planning further care and avoid miscommunication. Finally, equitable access to both the training and deployment of AI tools should be considered as the technology develops. Recently a study found severe disparity in geographic distribution of deep learning algorithms in clinical applications, with patient cohort data predominantly coming from three states: California, New York, and Massachusetts.[236] Representation is key, since patient outcomes and clinical care may vary across geography, ethnicity, and socioeconomic status. If these aspects are considered, modifications can be made to decrease the risk of health inequities such as race correction in clinical algorithms.[226]

The Future

The hype for AI is not new; in the history of artificial intelligence, there was a tremendous enthusiasm for AI that has led to several “AI winters” in the 1970’s and 1980’s, leading to periods of reduced funding and profound disillusionment. The current hype seems to mirror historical trends, particularly with claims of outperformance of deep learning tools versus clinicians in a recent systematic review.[241]

This time, however, things may be different. We have emerging clear frameworks to guide best practice and weigh claims of machine learning studies, there is an abundance of infrastructure for data storage, and the revolution of computational processing expands the capacity to handle ever-increasing amounts of data. More importantly, there is a multidisciplinary democratization of open-source machine learning tools through programming languages with pre-written, readily available software packages, such as TensorFlow, Scikit-Learn and PyTorch. The rapid emergence of a multidisciplinary approach and awareness of AI tools holds promise for the use of AI-assisted tools to guide and enhance our clinical practice.

Currently the role envisioned for machine learning tools is primarily to assist clinicians in making decisions for patient care. In the future, it is conceivable that integration of multimodal streams of data could frame the significance of findings, making decisions at a higher level than the clinician. When and if this occurs, the role for clinicians could change from gathering and analyzing information to spending time with patients helping them navigate their disease experience. Eric Topol of Scripps Research Institute, a leading thinker for AI and medicine, has emphasized the idea that when AI is able to deliver similar or superior results than humans, it still cannot replace the human side of empathetically being with patients.[252] The opportunity lies with reduced time burden in collating patient data and performing preliminary analyses, and instead spending the time interpreting results and managing patient expectations to help them cope as they progress through the treatment plan.

From optical biopsies and enhanced routine colonoscopies to selecting the optimal immunomodulator drug for inflammatory bowel disease, AI tools will potentially transform our practice by leveraging massive amounts of data to personalize care to the right patient, in the right amount, at the right time.

Key Challenges to EHR-based tools for Acute GIB

Challenges to electronic health record data are significant. EHR data is collected for clinical care, and not explicitly for research purposes. The data is reflective of “real world data” that may incorporate significant biases based on caregivers, patient population, and social context and is noisy, heterogeneous, and sparse.[91,253] In order to utilize machine learning appropriately and robustly for assisting management in acute gastrointestinal bleeding, one must consider issues under the broad framework of infrastructure and implementation.

The informatics infrastructure must have the capability for capturing, storing, and accessing data in a format that can be automatically extracted. This includes the type of electronic health record, as well as adequate data warehousing. There must be secure, adequate analytic platforms that can be used to deploy machine learning models in real time, which could include cloud computing. Finally, the healthcare system must have adequate expertise in algorithmic robustness. This includes understanding the risk of bias, interpretability, loss of performance over time, and impact of missingness and heterogeneity of EHR data.[254]

Implementation considerations are critical to understand how the tool fits into existing workflows and can be used to enhance care without introducing new burdens. Provider trust and usability is key, and it is important to consider principles of user-oriented design, qualitative evaluation of the intervention impact, and critical feedback. Furthermore, consideration of legal frameworks for liability and contingency plans for systemic errors are critical in ensuring patient safety. For example, if the machine learning score incorrectly classifies a patient as “very low risk” and the provider makes a decision using that information to discharge the patient, it is unclear what liability is attributable to the algorithm in the event of an adverse patient outcome.

Future Directions

The vision of an integrated machine learning-driven care pathway for patients with acute GIB begins with automated identification of patients using an NLP-based phenotype generated from a combination of text processing from provider notes and structured triage datafields. The identification would trigger a machine learning model to generate an initial risk assessment once the initial vital sign and laboratory measurements are made. At this point, the majority of very low risk patients at a high sensitivity threshold of >99% could be considered for outpatient management. Finally, for patients who are not very low risk, a dynamic risk assessment would estimate resuscitation needs for each 4–6-hour period to minimize organ damage from inadequate resuscitation and to optimize patients for endoscopic intervention. Implementation of this workflow into the electronic health record with a subsequent randomized controlled trial would be an important step to evaluate the efficacy of these systems in impacting the discharge rate of low-risk patients from the ED.

References

1. Peery AF, Crockett SD, Murphy CC, et al. Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2021. Gastroenterology 2021.

2. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380:1347-58.

3. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40-.

4. Gennatas ED, Friedman JH, Ungar LH, et al. Expert-augmented machine learning. Proc Natl Acad Sci U S A 2020:201906831.

5. Barkun AN, Almadi M, Kuipers EJ, et al. Management of Nonvariceal Upper Gastrointestinal Bleeding: Guideline Recommendations From the International Consensus Group. Annals of internal medicine 2019:10.7326/M19-1795.

6. Sung JJ, Chiu PW, Chan FKL, et al. Asia-Pacific working group consensus on non-variceal upper gastrointestinal bleeding: an update 2018. Gut 2018;67:1757-68.

7. Laine L, Jensen DM. Management of patients with ulcer bleeding. The American journal of gastroenterology 2012;107:345-60; quiz 61.

8. Strate LL, Gralnek IM. ACG Clinical Guideline: Management of Patients With Acute Lower Gastrointestinal Bleeding. Am J Gastroenterol 2016;111:459-74.

9. Gralnek IM, Dumonceau JM, Kuipers EJ, et al. Diagnosis and management of nonvariceal upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2015;47:a1-46.

10. Deshmukh F, Merchant SS. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. Official journal of the American College of Gastroenterology | ACG 9000;Publish Ahead of Print.

11. Stanley AJ, Ashley D, Dalton HR, et al. Outpatient management of patients with low-risk upper-gastrointestinal haemorrhage: multicentre validation and prospective evaluation. The Lancet 2009;373:42-7.

12. Banister T, Spiking J, Ayaru L. Discharge of patients with an acute upper gastrointestinal bleed from the emergency department using an extended Glasgow-Blatchford Score. BMJ Open Gastroenterol 2018;5:e000225.

13. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. Dig Dis Sci 2019.

14. Shung D, Au B, Taylor R, et al. Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding. Gastroenterology 2019.

15. Oakland K, Jairath V, Uberoi R, et al. Derivation and validation of a novel risk score for safe discharge after acute lower gastrointestinal bleeding: a modelling study. Lancet Gastroenterol Hepatol 2017;2:635-43.

16. Oakland K, Kothiwale S, Forehand T, et al. External Validation of the Oakland Score to Assess Safe Hospital Discharge Among Adult Patients With Acute Lower Gastrointestinal Bleeding in the US. JAMA Network Open 2020;3:e209630-e.

17. Das A, Ben-Menachem T, Cooper GS, et al. Prediction of outcome in acute lower-gastrointestinal haemorrhage based on an artificial neural network: internal and external validation of a predictive model. The Lancet 2003;362:1261-6.

18. Ayaru L, Ypsilantis P-P, Nanapragasam A, et al. Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting. PLOS ONE 2015;10:e0132485-e.

19. Loftus TJ, Brakenridge SC, Croft CA, et al. Neural network prediction of severe lower intestinal bleeding and the need for surgical intervention. J Surg Res 2017;212:42-7.

20. Liang PS, Saltzman JR. A National Survey on the Initial Management of Upper Gastrointestinal Bleeding. Journal of clinical gastroenterology 2014;48.

21. Hayes SM, Murray S, Dupuis M, Dawes M, Hawes IA, Barkun AN. Barriers to the Implementation of Practice Guidelines in Managing Patients with Nonvariceal Upper Gastrointestinal Bleeding: A Qualitative Approach. Canadian Journal of Gastroenterology 2010;24:878135.

22. Kosowicz RL, Strate LL. Predicting outcomes in lower gastrointestinal bleeding: more work ahead. Gastrointestinal endoscopy 2019;89:1014-6.

23. Shung D, Laine L. Machine Learning Prognostic Models for Gastrointestinal Bleeding Using Electronic Health Record Data. Am J Gastroenterol 2020;115:1199-200.

24. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24:198-208.

25. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annual Review of Biomedical Data Science 2018;1:53-68.

26. Peissig PL, Santos Costa V, Caldwell MD, et al. Relational machine learning for electronic health record-driven phenotyping. J Biomed Inform 2014;52:260-70.

27. Chen Y, Ghosh J, Bejan CA, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform 2015;55:82-93.

28. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:142-6.

29. Shung D, Tsay C, Laine L, et al. Early Identification of Patients with Acute Gastrointestinal Bleeding in the Emergency Department using Electronic Health Record Phenotyping. medRxiv 2020:2020.07.05.20136374.

30. Yahi A, Tatonetti NP. A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. AMIA Jt Summits Transl Sci Proc 2015;2015:64-8.

31. Kern EFO, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 2006;41:564-80.

32. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19:219-24.

33. Wilson FP, Shashaty M, Testani J, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. The Lancet 2015;385:1966-74.

34. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respiratory Research 2017;4:e000234.

35. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and Validation of an Electronic Health Record-Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. JAMA Netw Open 2018;1:e181018.

36. Jauk S, Kramer D, Großauer B, et al. Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study. Journal of the American Medical Informatics Association 2020;27:1383-92.

37. Laine L, Yang H, Chang SC, Datto C. Trends for incidence of hospitalization and death due to GI complications in the United States from 2001 to 2009. Am J Gastroenterol 2012;107:1190-5; quiz 6.

38. Abougergi MS, Travis AC, Saltzman JR. The in-hospital mortality rate for upper GI hemorrhage has decreased over 2 decades in the United States: a nationwide analysis. Gastrointest Endosc 2015;81:882-8 e1.

39. Wuerth BA, Rockey DC. Changing Epidemiology of Upper Gastrointestinal Hemorrhage in the Last Decade: A Nationwide Analysis. Dig Dis Sci 2018;63:1286-93.

40. Devani K, Radadiya D, Charilaou P, et al. Trends in hospitalization, mortality, and timing of colonoscopy in patients with acute lower gastrointestinal bleeding. Endosc Int Open 2021;9:E777-E89.

41. Wheat CL, Strate LL. Trends in Hospitalization for Diverticulitis and Diverticular Bleeding in the United States From 2000 to 2010. Clin Gastroenterol Hepatol 2016;14:96-103 e1.

42. Oakland K. Changing epidemiology and etiology of upper and lower gastrointestinal bleeding. Best practice & research Clinical gastroenterology 2019;42-43:101610.

43. [dataset] HCUP Nationwide Emergency Department Sample (NEDS). Healthcare Cost and Utilization Project (HCUP), 2006-2019. (Accessed November 2021, at https://www.hcup-us.ahrq.gov/nedsoverview.jsp.)

44. Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 Update and ICD-10 Translation. Am Health Drug Benefits 2019;12:188-97.

45. [dataset] Population Data. United States Census Bureau. (Accessed October 2021, at https://www.census.gov/topics/population/data.html.)

46. [dataset] HCUP Cost-to-Charge Ratio (CCR) Files for National Emergency Department Sample (NEDS). Healthcare Cost and Utilization Project (HCUP), 2012-2019. (Accessed December 2021, at https://www.hcup-us.ahrq.gov/db/ccr/ip-ccr/ip-ccr.jsp.)

47. [dataset] Medicare Inpatient Hospitals - by Geography and Service. Centers for Medicare & Medicaid Services. (Accessed December 2021, at https://data.cms.gov/provider-summary-by-type-of-service/medicare-inpatient-hospitals/medicare-inpatient-hospitals-by-geography-and-service.)

48. Dunn A, Grosse SD, Zuvekas SH. Adjusting Health Expenditures for Inflation: A Review of Measures for Health Services Research in the United States. Health Serv Res 2018;53:175-96.

49. [dataset] Gross Domestic Product. U.S. Bureau of Economic Analysis. (Accessed December 2021, at https://www.bea.gov/data/gdp/gross-domestic-product#gdp.)

50. Garthwaite C, Gross T, Notowidigdo M, Graves JA. Insurance Expansion and Hospital Emergency Department Access: Evidence From the Affordable Care Act. Ann Intern Med 2017;166:172-9.

51. Janke AT, Danagoulian S, Venkatesh AK, Levy PD. Medicaid expansion and resource utilization in the emergency department. Am J Emerg Med 2020;38:2586-90.

52. Taubman SL, Allen HL, Wright BJ, Baicker K, Finkelstein AN. Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment. Science 2014;343:263-8.

53. Sommers BD, Buchmueller T, Decker SL, Carey C, Kronick R. The Affordable Care Act has led to significant gains in health insurance and access to care for young adults. Health Aff (Millwood) 2013;32:165-74.

54. Crooks CJ, West J, Card TR. Comorbidities affect risk of nonvariceal upper gastrointestinal bleeding. Gastroenterology 2013;144:1384-93, 93 e1-2; quiz e18-9.

55. Hirode G, Saab S, Wong RJ. Trends in the Burden of Chronic Liver Disease Among Hospitalized US Adults. JAMA Netw Open 2020;3:e201997.

56. Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. Hepatology 2018;67:123-33.

57. Sorensen R, Hansen ML, Abildstrom SZ, et al. Risk of bleeding in patients with acute myocardial infarction treated with different combinations of aspirin, clopidogrel, and vitamin K antagonists in Denmark: a retrospective analysis of nationwide registry data. Lancet 2009;374:1967-74.

58. Abraham NS, Noseworthy PA, Inselman J, et al. Risk of Gastrointestinal Bleeding Increases With Combinations of Antithrombotic Agents and Patient Age. Clin Gastroenterol Hepatol 2020;18:337-46 e19.

59. Barkun AN, Bardou M, Kuipers EJ, et al. International consensus recommendations on the management of patients with nonvariceal upper gastrointestinal bleeding. Ann Intern Med 2010;152:101-13.

60. Villanueva C, Colomo A, Bosch A, et al. Transfusion strategies for acute upper gastrointestinal bleeding. N Engl J Med 2013;368:11-21.

61. Porter ME, Lee TH. From Volume to Value in Health Care: The Work Begins. JAMA 2016;316:1047-8.

62. Broderick RC, Fuchs HF, Harnsberger CR, et al. Increasing the Value of Healthcare: Improving Mortality While Reducing Cost in Bariatric Surgery. Obes Surg 2015;25:2231-8.

63. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27:759-69.

64. Cattaruzzi C, Troncon MG, Agostinis L, Garcia Rodriguez LA. Positive predictive value of ICD-9th codes for upper gastrointestinal bleeding and perforation in the Sistema Informativo Sanitario Regionale database. J Clin Epidemiol 1999;52:499-502.

65. Raiford DS, Perez Gutthann S, Garcia Rodriguez LA. Positive predictive value of ICD-9 codes in the identification of cases of complicated peptic ulcer disease in the Saskatchewan hospital automated database. Epidemiology 1996;7:101-4.

66. Cooper GS, Chak A, Lloyd LE, Yurchick PJ, Harper DL, Rosenthal GE. The accuracy of diagnosis and procedural codes for patients with upper GI hemorrhage. Gastrointest Endosc 2000;51:423-6.

67. Lanas A, García-Rodríguez LA, Polo-Tomás M, et al. Time trends and impact of upper and lower gastrointestinal bleeding and perforation in clinical practice. The American journal of gastroenterology 2009;104:1633-41.

68. Gerson LB, Fidler JL, Cave DR, Leighton JA. ACG Clinical Guideline: Diagnosis and Management of Small Bowel Bleeding. Am J Gastroenterol 2015;110:1265-87; quiz 88.

69. Peery AF, Crockett SD, Murphy CC, et al. Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2018. Gastroenterology 2019;156:254-72 e11.

70. Abougergi MS, Avila P, Saltzman JR. Impact of Insurance Status and Race on Outcomes in Nonvariceal Upper Gastrointestinal Hemorrhage: A Nationwide Analysis. J Clin Gastroenterol 2019;53:e12-e8.

71. Parast L, Mathews M, Martino S, Lehrman WG, Stark D, Elliott MN. Racial/Ethnic Differences in Emergency Department Utilization and Experience. J Gen Intern Med 2021.

72. American Medical Association and Association of American Medical Colleges. Advancing Health Equity: Guide on Language, Narrative and Concepts. 2021.

73. Khidir H, McWilliams JM, O'Malley AJ, Zaborski L, Landon BE, Smulowitz PB. Analysis of Consistency in Emergency Department Physician Variation in Propensity for Admission Across Patient Sociodemographic Groups. JAMA Netw Open 2021;4:e2125193.

74. Hall WJ, Chapman MV, Lee KM, et al. Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. Am J Public Health 2015;105:e60-76.

75. Agency for Healthcare Research and Quality. National Healthcare Quality and Disparities Report. Rockville, Maryland2021.

76. Rutter CM, May FP, Coronado GD, Pujol TA, Thomas EG, Cabreros I. Racism Is a Modifiable Risk Factor: Relationships Among Race, Ethnicity, and Colorectal Cancer Outcomes. Gastroenterology 2021.

77. Greenwood-Ericksen MB, Kocher K. Trends in Emergency Department Use by Rural and Urban Populations in the United States. JAMA Netw Open 2019;2:e191919.

78. Farm Labor. US Department of Agriculture, 2021. (Accessed January 10, 2021, at https://www.ers.usda.gov/topics/farm-economy/farm-labor/.)

79. James CV, Moonesinghe R, Wilson-Frederick SM, Hall JE, Penman-Aguilar A, Bouye K. Racial/Ethnic Health Disparities Among Rural Adults - United States, 2012-2015. MMWR Surveill Summ 2017;66:1-9.

80. Wollenman CS, Chason R, Reisch JS, Rockey DC. Impact of ethnicity in upper gastrointestinal hemorrhage. J Clin Gastroenterol 2014;48:343-50.

81. Eberly LA, Richterman A, Beckett AG, et al. Identification of Racial Inequities in Access to Specialized Inpatient Heart Failure Care at an Academic Medical Center. Circ Heart Fail 2019;12:e006214.

82. Laine L, Barkun AN, Saltzman JR, Martel M, Leontiadis GI. ACG Clinical Guideline: Upper Gastrointestinal and Ulcer Bleeding. The American journal of gastroenterology 2021;116:899-917.

83. Barkun AN, Almadi M, Kuipers EJ, et al. Management of Nonvariceal Upper Gastrointestinal Bleeding: Guideline Recommendations From the International Consensus Group. Annals of Internal Medicine 2019;171:805-22.

84. Stanley AJ, Ashley D, Dalton HR, et al. Outpatient management of patients with low-risk upper-gastrointestinal haemorrhage: multicentre validation and prospective evaluation. Lancet (London, England) 2009;373:42-7.

85. Sendak MP, Balu S, Schulman KA. Barriers to Achieving Economies of Scale in Analysis of EHR Data. A Cautionary Tale. Applied clinical informatics 2017;8:826-31.

86. Abougergi MS, Peluso H, Saltzman JR. Thirty-Day Readmission Among Patients With Non-Variceal Upper Gastrointestinal Hemorrhage and Effects on Outcomes. Gastroenterology 2018;155:38-46.e1.

87. Campbell HE, Stokes EA, Bargo D, et al. Costs and quality of life associated with acute upper gastrointestinal bleeding in the UK: cohort analysis of patients in a cluster randomised trial. BMJ Open 2015;5:e007230.

88. Oakland K, Chadwick G, East JE, et al. Diagnosis and management of acute lower gastrointestinal bleeding: guidelines from the British Society of Gastroenterology. Gut 2019;68:776-89.

89. Laine L. Risk Assessment Tools for Gastrointestinal Bleeding. Clinical Gastroenterology and Hepatology 2016;14:1571-3.

90. Laursen SB, Oakland K, Laine L, et al. ABC score: a new risk score that accurately predicts mortality in acute upper and lower gastrointestinal bleeding: an international multicentre study. Gut 2020:gutjnl-2019-320002.

91. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395-405.

92. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144-51.

93. Deshmukh F, Merchant SS. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. The American journal of gastroenterology 2020;115:1657-68.

94. Levi R, Carli F, Arévalo AR, et al. Artificial intelligence-based prediction of transfusion in the intensive care unit in patients with gastrointestinal bleeding. BMJ Health Care Inform 2021;28:e100245.

95. Barkun AN, Bardou M, Kuipers EJ, et al. International consensus recommendations on the management of patients with nonvariceal upper gastrointestinal bleeding. Ann Intern Med 2010;152:101-13.

96. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association : JAMIA 2013;20:e206-e11.

97. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. Journal of the American Medical Informatics Association : JAMIA 2013;20:e226-e31.

98. Angus DC. Fusing Randomized Trials With Big Data: The Key to Self-learning Health Care Systems?Fusing Randomized Trials With Big DataFusing Randomized Trials With Big Data. JAMA 2015;314:767-8.

99. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885.

100. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. arXiv e-prints2019:arXiv:1902.07669.

101. Wang Y, Liu S, Afzal N, et al. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. arXiv e-prints2018:arXiv:1802.00400.

102. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One 2019;14:e0212778.

103. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018:arXiv:1810.04805.

104. Chang D, Balazevic I, Allen C, Chawla D, Brandt C, Taylor RA. Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. 2020:arXiv:2006.13774.

105. Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis 2013;19:1411-20.

106. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010;62:1120-7.

107. Xia Z, Secor E, Chibnik LB, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS One 2013;8:e78927.

108. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol 2019;34:557-65.

109. Wilkinson T, Ly A, Schnier C, et al. Identifying dementia cases with routinely collected health data: A systematic review. Alzheimers Dement 2018;14:1038-51.

110. Rubbo B, Fitzpatrick NK, Denaxas S, et al. Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. International Journal of Cardiology 2015;187:705-11.

111. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. Journal of the American Medical Informatics Association 2018;25:1618-25.

112. Elkin PL, Ruggieri AP, Brown SH, et al. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. Proc AMIA Symp 2001:159-63.

113. Medicine Io. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: The National Academies Press; 2011.

114. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv Res 2021;21:200.

115. Hipp R, Abel E, Weber RJ. A Primer on Clinical Pathways. Hosp Pharm 2016;51:416-21.

116. Coffey RJ, Richards JS, Remmert CS, LeRoy SS, Schoville RR, Baldwin PJ. An introduction to critical paths. Qual Manag Health Care 1992;1:45-54.

117. Damiani G, Pinnarelli L, Colosimo SC, et al. The effectiveness of computerized clinical guidelines in the process of care: a systematic review. BMC Health Serv Res 2010;10:2-.

118. Campbell R. The five "rights" of clinical decision support. J ahima 2013;84:42-7; quiz 8.

119. Peery AF, Crockett SD, Murphy CC, et al. Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2018. Gastroenterology 2019;156:254-72.e11.

120. Hearnshaw SA, Logan RF, Lowe D, Travis SP, Murphy MF, Palmer KR. Acute upper gastrointestinal bleeding in the UK: patient characteristics, diagnoses and outcomes in the 2007 UK audit. Gut 2011;60:1327-35.

121. Strate LL, Ayanian JZ, Kotler G, Syngal S. Risk factors for mortality in lower intestinal bleeding. Clin Gastroenterol Hepatol 2008;6:1004-10; quiz 955-.

122. Villanueva C, Colomo A, Bosch A, et al. Transfusion Strategies for Acute Upper Gastrointestinal Bleeding. New England Journal of Medicine 2013;368:11-21.

123. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for uppergastrointestinal haemorrhage. The Lancet 2000;356:1318-21.

124. Shung DL, Au B, Taylor RA, et al. Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding. Gastroenterology 2019.

125. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. Dig Dis Sci 2019;64:2078-87.

126. Herasevich V, Pieper MS, Pulido J, Gajic O. Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation. Journal of the American Medical Informatics Association 2011;18:639-44.

127. Umscheid CA, Betesh J, VanZandbergen C, et al. Development, implementation, and impact of an automated early warning and response system for sepsis. J Hosp Med 2015;10:26-31.

128. Mohamadlou H, Lynn-Palevsky A, Barton C, et al. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. Canadian Journal of Kidney Health and Disease 2018;5:2054358118776326.

129. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. Annals of Surgery 2019;269.

130. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116-9.

131. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine 2015;7:299ra122.

132. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1:18.

133. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101:E215-20.

134. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific Data 2016;3:160035.

135. Simonov M, Ugwuowo U, Moreira E, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: A descriptive modeling study. PLoS Med 2019;16:e1002861.

136. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1-22.

137. Lin Y-W, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PloS one 2019;14:e0218942-e.

138. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

139. Oakland K, Chadwick G, East JE, et al. Diagnosis and management of acute lower gastrointestinal bleeding: guidelines from the British Society of Gastroenterology. Gut 2019;68:776.

140. Laine L, Jensen DM. Management of Patients With Ulcer Bleeding. American Journal of Gastroenterology 2012;107.

141. Sung JJY, Chiu PWY, Chan FKL, et al. Asia-Pacific working group consensus on non-variceal upper gastrointestinal bleeding: an update 2018. Gut 2018;67:1757.

142. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 2005;18:602-10.

143. Felix A. Gers JS, and Fred Cummins. Learning to forget: Continual prediction with LSTM. ICANN Ninth International Conference on Artificial Neural Networks1999:850-5.

144. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-80.

145. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association 2016;24:198-208.

146. Baradarian R, Ramdhaney S, Chapalamadugu R, et al. Early intensive resuscitation of patients with upper gastrointestinal bleeding decreases mortality. Am J Gastroenterol 2004;99:619-22.

147. Oakland K, Guy R, Uberoi R, et al. Acute lower GI bleeding in the UK: patient characteristics, interventions and outcomes in the first nationwide audit. Gut 2018;67:654.

148. Kizzier-Carnahan V, Artis KA, Mohan V, Gold JA. Frequency of Passive EHR Alerts in the ICU: Another Form of Alert Fatigue? J Patient Saf 2019;15:246-50.

149. Guo T, Lin T, Antulov-Fantulin N. Exploring Interpretable LSTM Neural Networks over Multi-Variable Data. 2019:arXiv:1905.12034.

150. Fan F, Xiong J, Li M, Wang G. On Interpretability of Artificial Neural Networks: A Survey. 2020:arXiv:2001.02522.

151. Gu J, Tresp V. Contextual Prediction Difference Analysis for Explaining Individual Image Classifications. 2019:arXiv:1910.09086.

152. Fong R, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. 2017:arXiv:1704.03296.

153. Tulio Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv e-prints2016.

154. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine 2019;17:195.

155. Köpcke F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013;13:37.

156. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J Biomed Inform 2017;68:112-20.

157. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. Diagnostic and Prognostic Research 2020;4:8.

158. van der Maaten LJP, van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. Journal of machine learning research 2008;9:2579-605.

159. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018:arXiv:1802.03426.

160. Coifman RR, Lafon S. Diffusion maps. Applied and Computational Harmonic Analysis 2006;21:5-30.

161. Moon KR, van Dijk D, Wang Z, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2019;37:1482-92.

162. Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an algorithm. In: Dietterich T, Becker S, Ghahramani Z, editors. Advances in Neural Information Processing Systems; 2002: MIT Press.

163. Van Der Maaten L, Postma E, Van den Herik J. Dimensionality reduction: a comparative. J Mach Learn Res 2009;10:13.

164. Criminisi A, Shotton J, Konukoglu E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Found Trends Comput Graph Vis 2012;7:81–227-81–.

165. Moon KR, Dijk Dv, Wang Z, et al. Visualizing Structure and Transitions for Biological Data Exploration. bioRxiv 2018:120378.

166. Izenman AJ. Introduction to manifold learning. Wiley Interdisciplinary Reviews: Computational Statistics 2012;4:439-46.

167. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees: Routledge; 2017.

168. Tsallis C. Possible generalization of Boltzmann-Gibbs statistics. Journal of statistical physics 1988;52:479-87.

169. Hegde C, Wakin M, Baraniuk R. Random projections for manifold learning. Advances in neural information processing systems 2007;20:641-8.

170. Dasgupta S, Freund Y. Random projection trees and low dimensional manifolds. Proceedings of the fortieth annual ACM symposium on Theory of computing; 2008. p. 537-46.

171. Verma N, Kpotufe S, Dasgupta S. Which spatial partition trees are adaptive to intrinsic dimension? arXiv preprint arXiv:12052609 2012.

172. McCartin-Lim M, McGregor A, Wang R. Approximate principal direction trees. arXiv preprint arXiv:12064668 2012.

173. Madhyastha M, Li G, Strnadová-Neeley V, et al. Geodesic Forests. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020. p. 513-23.

174. Peyré G, Cuturi M. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning 2019;11:355-607.

175. Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 2013;26:2292-300.

176. Le T, Yamada M, Fukumizu K, Cuturi M. Tree-sliced variants of wasserstein distances. arXiv preprint arXiv:190200342 2019.

177. Devine EB, Capurro D, van Eaton E, et al. Preparing electronic clinical data for quality improvement and comparative effectiveness research: the SCOAP CERTAIN automation and validation project. EGEMS 2013;1.

178. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. JAMA Network Open 2021;4:e210184-e.

179. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American journal of epidemiology 1995;142:1255-64.

180. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research 2007;16:219-42.

181. Little RJ. A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association 1988;83:1198-202.

182. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation 2000;101:e215-e20.

183. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific data 2018;5:1-13.

184. Stanley AJ, Laine L, Dalton HR, et al. Comparison of risk scoring systems for patients presenting with upper gastrointestinal bleeding: international multicentre prospective study. BMJ (Clinical research ed) 2017;356.

185. Cosgriff CV, Celi LA, Ko S, et al. Developing well-calibrated illness severity scores for decision support in the critically ill. npj Digital Medicine 2019;2:76.

186. Backurs A, Dong Y, Indyk P, Razenshteyn I, Wagner T. Scalable nearest neighbor search for optimal transport. International Conference on Machine Learning; 2020: PMLR. p. 497-506.

187. Bojchevski A, Günnemann S. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. arXiv preprint arXiv:170703815 2017.

188. Schulz S, Klein GO. SNOMED CT – advances in concept mapping, retrieval, and ontological foundations. Selected contributions to the Semantic Mining Conference on SNOMED CT (SMCS 2006). BMC Medical Informatics and Decision Making 2008;8:S1.

189. Schulz S, Klein GO. SNOMED CT–advances in concept mapping, retrieval, and ontological foundations. Selected contributions to the Semantic Mining Conference on SNOMED CT (SMCS 2006). Springer; 2008:1-3.

190. Indyk P. Fast color image retrieval via embeddings. Workshop on Statistical and Computational Theories of Vision (at ICCV), 2003; 2003.

191. Tong AY, Huguet G, Natik A, et al. Diffusion Earth Mover's Distance and Distribution Embeddings. International Conference on Machine Learning; 2021: PMLR. p. 10336-46.

192. Indyk P, Thaper N. Fast image retrieval via embeddings. 3rd international workshop on statistical and computational theories of vision; 2003. p. 5.

193. Shirdhonkar S, Jacobs DW. Approximate earth mover's distance in linear time. 2008 IEEE Conference on Computer Vision and Pattern Recognition; 2008: IEEE. p. 1-8.

194. Gavish M, Nadler B, Coifman RR. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. ICML; 2010.

195. Huizing G-J, Peyré G, Cantini L. Optimal Transport improves cell-cell similarity inference in single-cell omics data. bioRxiv 2021.

196. Bellazzi R, Codegoni A, Gualandi S, Nicora G, Vercesi E. The Gene Mover's Distance: Single-cell similarity via Optimal Transport. arXiv preprint arXiv:210201218 2021.

197. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. International conference on machine learning; 2017: PMLR. p. 214-23.

198. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. Advances in neural information processing systems 2017;30.

199. Tong A, Wolf G, Krishnaswamy S. Fixing bias in reconstruction-based anomaly detection with lipschitz discriminators. Journal of Signal Processing Systems 2021:1-15.

200. Le T, Yamada M, Fukumizu K, Cuturi M. Tree-sliced variants of Wasserstein distances. Advances in neural information processing systems 2019;32.

201. Balaji Y, Chellappa R, Feizi S. Robust optimal transport with applications in generative modeling and domain adaptation. Advances in Neural Information Processing Systems 2020;33:12934-44.

202. Chizat L, Peyré G, Schmitzer B, Vialard F-X. Unbalanced optimal transport: Dynamic and Kantorovich formulations. Journal of Functional Analysis 2018;274:3090-123.

203. Liero M, Mielke A, Savaré G. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. Inventiones mathematicae 2018;211:969-1117.

204. Fatras K, Séjourné T, Flamary R, Courty N. Unbalanced minibatch optimal transport; applications to domain adaptation. International Conference on Machine Learning; 2021: PMLR. p. 3186-97.

205. Genevay A, Peyré G, Cuturi M. Learning generative models with sinkhorn divergences. International Conference on Artificial Intelligence and Statistics; 2018: PMLR. p. 1608-17.

206. Fatras K, Zine Y, Flamary R, Gribonval R, Courty N. Learning with minibatch Wasserstein: asymptotic and gradient properties. arXiv preprint arXiv:191004091 2019.

207. Caffarelli LA, McCann RJ. Free boundaries in optimal transport and Monge-Ampere obstacle problems. Annals of mathematics 2010:673-730.

208. Pele O, Werman M. Fast and robust earth mover's distances. 2009 IEEE 12th international conference on computer vision; 2009: IEEE. p. 460-7.

209. Mukherjee D, Guha A, Solomon JM, Sun Y, Yurochkin M. Outlier-robust optimal transport. International Conference on Machine Learning; 2021: PMLR. p. 7850-60.

210. Coifman RR, Maggioni M. Diffusion wavelets. Applied and computational harmonic analysis 2006;21:53-94.

211. Leeb W, Coifman R. Hölder–Lipschitz norms and their duals on spaces with semigroups, with applications to earth mover's distance. Journal of Fourier Analysis and Applications 2016;22:910-53.

212. Roski J, Maier EJ, Vigilante K, Kane EA, Matheny ME. Enhancing trust in AI through industry self-governance. J Am Med Inform Assoc 2021;28:1582-90.

213. Vokinger KN, Gasser U. Regulating AI in medicine in the United States and Europe. Nat Mach Intell 2021;3:738-9.

214. Li Q, Lan L, Zeng N, et al. A Framework for Big Data Governance to Advance RHINs: A Case Study of China. IEEE Access 2019;7:50330-8.

215. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. J Public Health Policy 2021;42:602-11.

216. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med 2018;178:1544-7.

217. Johnson-Mann CN, Loftus TJ, Bihorac A. Equity and Artificial Intelligence in Surgical Care. JAMA Surg 2021;156:509-10.

218. Haider AH, Scott VK, Rehman KA, et al. Racial disparities in surgical care and outcomes in the United States: a comprehensive review of patient, provider, and systemic factors. J Am Coll Surg 2013;216:482-92.e12.

219. Arpey NC, Gaglioti AH, Rosenbaum ME. How Socioeconomic Status Affects Patient Perceptions of Health Care: A Qualitative Study. J Prim Care Community Health 2017;8:169-75.

220. Ng JH, Ye F, Ward LM, Haffer SC, Scholle SH. Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. Health Aff (Millwood) 2017;36:548-52.

221. Greenwood-Ericksen MB, Kocher K. Trends in Emergency Department Use by Rural and Urban Populations in the United States. JAMA Network Open 2019;2:e191919-e.

222. Straw I, Callison-Burch C. Artificial Intelligence in mental health and the biases of language based models. PLoS One 2020;15:e0240376.

223. Becker D, Lamb S. Sex bias in the diagnosis of borderline personality disorder and posttraumatic stress disorder. Professional Psychology: Research and Practice 1994;25:55-61.

224. Jane JS, Oltmanns TF, South SC, Turkheimer E. Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. J Abnorm Psychol 2007;116:166-75.

225. Snowden LR. Bias in mental health assessment and intervention: theory and evidence. Am J Public Health 2003;93:239-43.

226. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine 2020;383:874-82.

227. Western MJ, Armstrong MEG, Islam I, Morgan K, Jones UF, Kelson MJ. The effectiveness of digital interventions for increasing physical activity in individuals of low socioeconomic status: a systematic review and meta-analysis. International Journal of Behavioral Nutrition and Physical Activity 2021;18:148.

228. Colvonen PJ, DeYoung PN, Bosompra N-OA, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. Sleep 2020;43:zsaa159.

229. Martinez-Martin N, Luo Z, Kaushal A, et al. Ethical issues in using ambient intelligence in health-care settings. The Lancet Digital Health 2021;3:e115-e23.

230. Sikstrom L, Maslej MM, Hui K, Findlay Z, Buchman DZ, Hill SL. Conceptualising fairness: three pillars for medical algorithms and health equity. BMJ Health Care Inform 2022;29.

231. Wei J, Mehta PK, Grey E, et al. Sex-based differences in quality of care and outcomes in a health system using a standardized STEMI protocol. Am Heart J 2017;191:30-6.

232. Hao Y, Liu J, Liu J, et al. Sex Differences in In-Hospital Management and Outcomes of Patients With Acute Coronary Syndrome. Circulation 2019;139:1776-85.

233. MacDorman MF, Thoma M, Declcerq E, Howell EA. Racial and Ethnic Disparities in Maternal Mortality in the United States Using Enhanced Vital Records, 2016–2017. Am J Public Health 2021;111:1673-81.

234. Oala L, Murchison AG, Balachandran P, et al. Machine Learning for Health: Algorithm Auditing & Quality Control. J Med Syst 2021;45:105.

235. Samorani M, Harris SL, Blount LG, Lu H, Santoro MA. Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. Manufacturing & Service Operations Management 2021.

236. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA 2020;324:1212-3.

237. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. JAMA 2020;324:1397-8.

238. Embi PJ. Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence–Driven Health Care for Effectiveness and Equity. JAMA Network Open 2021;4:e214622-e.

239. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58.

240. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. New England Journal of Medicine 2019;380:1347-58.

241. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. Bmj 2020;368:m689.

242. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial Intelligence and the Implementation Challenge. J Med Internet Res 2019;21:e13659.

243. Sendak MP, Ratliff W, Sarro D, et al. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. JMIR Med Inform 2020;8:e15182.

244. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. npj Digital Medicine 2020;3:107.

245. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020;370:m3210.

246. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164.

247. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. JAMA 2020.

248. Mori Y, Kudo S-e, East JE, et al. Cost savings in colonoscopy with artificial intelligence-aided polyp diagnosis: an add-on analysis of a clinical trial (with&#xa0;video)<sup>&#x2217;</sup>. Gastrointestinal Endoscopy 2020;92:905-11.e1.

249. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287-9.

250. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30-6.

251. Shen FX, Wolf SM, Gonzalez RG, Garwood M. Ethical Issues Posed by Field Research Using Highly Portable and Cloud-Enabled Neuroimaging. Neuron 2020;105:771-5.

252. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again: Basic Books, Inc.; 2019.

253. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144-51.

254. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. JMIR Med Inform 2018;6:e11-e.