Adapting Trauma Outcome Prediction Models to Individual Facilities using Transfer Learning

Carly Marie Eckert

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Stephen Mooney, Chair

Brianna Mills

Randall Burd

Program Authorized to Offer Degree:

Department of Epidemiology

University of Washington

Abstract

Adapting Trauma Outcome Prediction Models to Individual Facilities using Transfer Learning

Carly Marie Eckert

Chair of the Supervisory Committee:
Steve Mooney
Department of Epidemiology

With the increasing availability of big data and advanced computational techniques, machine learning (ML) models are becoming common in medicine and healthcare. Generalizable models, models that can be applied to any setting or patient cohort, are described as a goal of ML, yet sacrifices in performance are required to demonstrate such broad applicability. To date, there is a gap in the use of modeling techniques that can learn overarching patterns from larger data sets that, when incorporated with local data, can better model facility-specific trends. Transfer learning (TL) techniques incorporate two disparate data sets, a source and a target. TL techniques learn patterns from the initial data set (**the source**) and apply relevant knowledge to the modeling task of a second data set (**the target**). We evaluated the use of TL in trauma outcomes prediction modeling at the level of the individual hospital to assess the impact of this approach when using small record sets. We considered two feature set variations when developing our logistic regression predictive models: the shared feature set between the source and the target and the union of the two feature sets. Compared to baseline approaches, TL-based models did not result in consistent improvements in predictive model performance.

Table of Contents

Primary Manuscript 1

Evaluation of Facility Specific Predictive Models with Transfer Learning

Abstract

Importance:

Small record sets are common in healthcare and using them in machine learning (ML) problems is challenging as these methods require large amounts of data. These challenges are compounded when rare outcomes are of interest as examples are infrequent. Transfer learning (TL) provides an approach to augment datasets limited by small record sets with additional examples for ML models to learn while retaining the local context of the original data.

Objective:

We evaluated the efficacy of TL in predicting inpatient death following traumatic injury by augmenting hospital-specific datasets from the National Trauma Data Bank (NTDB) with another publicly available dataset (eICU-CRD). We compared facility-specific TL models with three baseline models built using traditional ML approaches to characterize the contexts in which TL may lead to improved model performance.

Study Design / Setting/ Participants:

Trauma encounters between 2014 and 2015 constituted the training data. From the eICU dataset, encounters were selected based on the diagnosis of acute injury (n=6,720). From the NTDB, thirty Level II trauma facilities were selected by stratified random sampling. All encounters within the facilities that received ICU care during the encounter were included (n=16,140). The TL model was limited to the features shared across the two datasets. The baseline models were built using only facility-specific NTDB data and included an 'all features' model, a 'shared features' model, and a model based on the New Trauma Score. Facility-specific models (n=30) were built using each approach and each model, including the TL model, was tested on NTDB facility-specific data from 2016.

Results:

The TL model exceeded the performance of the best-performing baseline for 12/30 facilities. However, when comparing approaches limited to the shared feature space, the 'shared features' baseline and the TL approach, the TL approach tended to outperform the traditional ML approach (n=24). The 'all features' baseline model performed best in 13/30 facilities. TL performed better than other approaches for facilities in which transfer rates were low (<10%). TL did not perform well when patient transfer rates exceeded 30%.

Conclusions:

TL methods promise to borrow information from larger source datasets to fit models to record sets too small for conventional ML. Based on our logistic regression approach, a small number of records in the target dataset is an insufficient characterization of scenarios when TL is useful, as source and target alignment within the feature space affects performance. TL methods did not result in substantial improvements in predictive performance across facilities.

Introduction

AI tools and systems are increasingly used in healthcare settings.[1] These tools, including prognostic models built with machine learning (ML) algorithms, usually require large datasets to perform well at their prescribed task.[2] It is hypothesized that when datasets are large and varied 'enough', generalizable models are possible. Generalizable models perform their task within a different context from which they are trained, known as external validity.[3] Unfortunately, models developed on large datasets often fail to perform well in specific settings, and the promise of generalizable models remains unrealized.[4,5] Local context, particularly in healthcare where facility-specific processes and patient populations may be variable, is vital to adequately capture data patterns.[5] Locally developed models can incorporate key variables that describe the processes, facility characteristics, and operations that may be specific to one facility or distinguish one facility from another.[5] Additionally, the development and application of local models may reveal opportunities to improve upon hospital-specific factors and provide actionable insights for users of the model, such as identifying changes to order forms that increase guideline adherence in critical care management, thereby realizing the goals of a learning health system.[6] However, individual hospitals may not have sufficient data to understand the patterns they are interested in exploring. Without sufficient data, prognostic models cannot accurately predict the outcome of interest due to overfitting and will be unable to generalize.[7]

This limitation of insufficient data results from small datasets, a challenging and ill-defined problem in the era of 'big data'.[8] Small datasets are common in healthcare where relevant patient records may number in the dozens to hundreds and where simply obtaining more data is impracticable.[9] Small datasets are challenging for ML researchers as ML algorithms must have adequate examples from which to learn.[10] ML algorithms work by identifying patterns in datasets and associating the patterns to the outcome. A larger dataset provides the algorithm with more examples of the problem being studied from which the algorithm can learn. In addition to the size of the dataset, the number of samples representing each outcome is crucial and may similarly be constrained with small datasets. When only a few examples of one outcome are in the data, the dataset is considered imbalanced. With imbalanced datasets, there is a substantial difference in the prevalence of the rare event (typically the positive class) compared to the prevalence of instances where the event does not occur (the 'majority class', typically the negative class).[11] When small datasets are also imbalanced, an ML algorithm often does not have sufficient examples to learn the patterns of features associated with the rare outcome. Consequently, the algorithm misclassifies the

positive class more frequently.[12] The positive class may be of greater importance clinically as it generally indicates a state of patient need (e.g., heart attack, readmission, missed clinical appointment). Addressing imbalanced datasets due to rare outcomes is a well-researched problem in ML, and solutions include undersampling the majority class or oversampling the minority class to achieve balance.[13] When small datasets are highly imbalanced, the problem is termed "absolute rarity" and remains an open question in ML research.[14] With absolute rarity, an ML model needs more records to train on and greater representativeness of the positive class.

One approach to address the problem of small record sets involves using transfer learning (TL). TL is a computational technique within ML that applies knowledge from one task to another.[15] TL techniques learn patterns from a larger dataset (**the source**) and apply relevant knowledge to the modeling task of a second, usually smaller, dataset (**the target**). Specifically, when applying TL methods, the source dataset is incorporated into the target data to provide additional examples of training data from which the model can learn. If the source dataset is sufficiently similar to the target data, augmenting the target with the source dataset will improve the discriminatory capability of the ML model compared to a target-only model.[15] Thus, TL methods use source data to better understand the general mechanism of interest and use the target data to finely tune that mechanism on local data. In this way, TL offers an approach that acknowledges the necessity of local data to capture variability across health systems while alleviating the problem of insufficiently sized record sets. Due to the hierarchical structure of care delivery, trauma and injury care may benefit sharply from leveraging TL to build better prognostic models and will be the context for this work.

One challenge for TL is avoiding negative transfer, defined as reduced performance metrics compared to using the target data alone.[15,16,17] With negative transfer the general mechanism of interest between the source and target datasets are too disparate to realize any performance advantage in combining the datasets. To avoid negative transfer, the source dataset must be aligned or related to the target data, however, the degree of relatedness is unspecified. In surveys of transfer learning, researchers refer to alignment in the domain of study, the feature spaces involved, the distribution of independent predictors (marginal probabilities), and the likelihood of the outcome given a distribution of predictors (conditional probability).[15,16] Notably, both general feature concepts and precise definitions of the features must be considered; otherwise, context feature bias can occur.[16] Homogeneous TL methods, a specific approach to TL, can be applied when the feature spaces between the source and target datasets are identical.[16]

Despite this concern regarding negative transfer, the use of TL is becoming more prevalent in medical literature, particularly when used with convolutional neural networks for image classification, such as skin cancer diagnosis and radiographic imaging.[18] Other examples of TL methods in healthcare research include those of Lee *et al.,* who used TL methods to assess mortality following surgical procedures and highlighted the deficiencies of models built on large registries in predicting outcomes among patients at small hospitals.[19] Wiens *et al.* used TL methods to predict patient infection with *Clostridium difficile* while incorporating data from multiple hospital EHRs.[20] Desautels *et al.* used TL to predict the likelihood of ICU readmission and inpatient death using data extracted from a large urban hospital and ICU research database.[21] Curth *et al.* demonstrated different domain adaptation methods for predicting the likelihood of patient readmission after ICU discharge.[22] In each of these examples, the authors illustrated the benefit of TL with enhanced performance in the prediction task as defined by improved classification metrics. The examples provided by *Lee et al.* and *Desautels et al.* used widely available source datasets to augment their training data while the work of *Curth et al.* and *Wiens et al.* combined data from different hospitals included in their study to compose their source datasets.

This study used a homogeneous TL approach to develop and evaluate local models predicting mortality following traumatic injury at Level II trauma facilities. By predicting inpatient death following injury among this subset of patient encounters, we evaluated the efficacy of TL in addressing the challenges of rare outcomes and small datasets. We compared facility-specific models generated with TL to baseline models generated using only target-specific data. We described situations where this approach may be effective and characterized the facilities where TL was advantageous over traditional ML modeling. We hypothesized that models developed using TL methods would perform better than those created with target-only facility-specific data among small facilities and among facilities with feature distributions similar to those of the source data.

Methods

1. *Data Sources*

Source Data:

The eICU Collaborative Research Database (eICU) is a de-identified electronic health record (EHRs) extract from patients admitted to intensive care units (ICU) at 208 facilities across the United States.[23] The data contained in the eICU database were collected during the

course of care of patients receiving critical care services via a telehealth and telemonitoring platform. Encounters from the eICU database number over 200,000 and represent encounters from 2014-2015. This database is publicly available with approval for research purposes. This database includes encounter-level data on patient demographics, admission diagnoses, vital sign measurements, laboratory results, and treatment plans.

Target Data:

The National Trauma Data Bank (NTDB) was developed by the American College of Surgeons (ACS) and is the largest aggregation of trauma registry data in the United States. The NTDB collects and aggregates data from US hospitals and then creates and distributes datasets for research purposes.[24] The NTDB collects data such as demographics, injury severity, and physiologic data from trauma encounters. In the NTDB, we identified patients who received care at ACS certified Level II trauma facilities between 2014 and 2016. Level II trauma centers manage acutely ill patients providing definitive care when needed. The University of Washington Institutional Review Board determined that this study was exempt from IRB review because it did not meet the criteria for human subject research.

Based on the defining characteristics of these data sources, we do not expect there to be overlap in included patient encounters between them as the use of telehealth care in level II trauma centers is not common practice.

2. Study Cohort

Source Data:

We included all patient encounters in the eICU dataset for evaluation. The cohort was defined based on patient encounters associated with traumatic injuries. ICD-9 CM and ICD-10 CM coding were inconsistently available; therefore, string matching of the word “trauma” in the diagnosis was used to select the cohort (Figure 1). The excluded diagnoses for non-acute trauma-related admissions are detailed in *Supplement: eICU Cohort Determination.*

Target Data:

Thirty Level II trauma facilities and their associated patient care encounters were selected from the NTDB, with the only criterion being that the facility persisted in the NTDB across the years of study inclusion (2014 – 2016). The facilities were selected via stratified randomized sampling based on the bed size of the facility, with 60% of the facilities (n=18)

having <400 patient beds and 40% of the facilities (n=12) having ≥ 400 patient beds, with a sampling ratio equivalent to the prevalence of large and small Level II facilities in the NTDB. Encounters from 2014 and 2015 where intensive care was required were included in the evaluation (Figure 2). Data from 2016 trauma encounters were retained for facility-specific model testing. Facility identifiers for each encounter were maintained so facility-specific data could be grouped and evaluated throughout the study. Data after 2016 was excluded from evaluation as consistent facility identifiers were no longer made public via the NTDB.

3. *Outcome & Predictors*

Source Data:

The primary endpoint was death during the inpatient encounter. eICU encounters without a discharge disposition were excluded from the database. All eICU tables containing structured data elements were considered for inclusion, with the caveat that these fields or their equivalent must also be available in the NTDB. Demographic information, comorbidities, vital signs, and other variables were included (Table 1). Injury severity is a known predictor of mortality from trauma, and diagnostic coding is commonly used to assess the severity of injury, often in the form of the Injury Severity Score (ISS).[25] The ISS is an aggregate numerical score that reflects the three most severely injured body regions. Although the ISS is available in the NTDB, it is unavailable in the eICU dataset. This value could not be determined because the diagnosis codes (ICD-9-CM) did not translate to Abbreviated Injury Scale (AIS) scores, the component scores that combine to generate ISS, in nearly 60% of the codes used.[26] Instead, proxies for injury severity included the following features: Glasgow Coma Scale (GCS) total and component scores (motor, verbal, and eye opening), blood transfusion requirement, patient ventilator requirement, and the nature of injury descriptors. The care plan tables were evaluated for other consistently captured data that could be used as proxies for injury severity. The first vital sign measurements and GCS scores were used to ensure alignment with the vital signs and GCS captured in the NTDB.

Information on advance directives, including *do not resuscitate* orders, was available and used to formulate a binary feature. We generated a binary feature to reflect a patient's ventilator requirements. A binary feature was generated to indicate the presence or absence of any comorbidity based on the patient's medical history. We generated this feature for both the source and target data from the supplied list of patient comorbidities in each dataset, the details of which are provided in *Supplement: Feature Development.* We limited our features to the information available within 24 hours of patient admission. Timestamps were removed from

eICU data; however, the data included time offsets based on the time of patient admission. Data collected after an offset of 1440 (minutes within a 24-hour time period) were excluded from the analysis.

Target Data:

Death and discharge to hospice care were used to define the primary endpoint.[27] NTDB encounters that did not have a discharge disposition were excluded. Any variables included in the TL model from the NTDB required matching variables from the eICU dataset. This requirement eliminated many variables from the NTDB for consideration, such as the ISS, type of injury, intent of injury, and mechanism of injury. The distribution of these variables within the training and test populations is described in Table 2. We retained the nature of injury data as we were able to devise a corollary feature in the source data. For the homogeneous TL approach, we needed to ensure that the features reflected similar patient occurrences across the datasets as much as possible.[16] Field-collected vital signs and GCS scores were excluded, and emergency room vital signs and GCS scores were used. Features were generated from the available variables, including ventilator requirements, blood transfusion requirements, and comorbidity indicators (see *Supplement: Feature Development* for details on feature construction).

Across both source and target datasets, continuous predictors were evaluated for outliers or nonsensical values. The NTDB uses -1 and -2 to denote not applicable and unknown values, respectively.[28] We treated both indicator values as missing (details in the *Supplement: Missing Data*). Continuous predictors were scaled during the modeling and evaluation processes. These included age, systolic blood pressure, pulse rate, and oxygen saturation. GCS scores were encoded as integers but were allowed to function as continuous variables when scaled. The missingness of the potential predictors was evaluated during the facility modeling process after the NTDB data were combined with the eICU data. We evaluated missingness at the facility level to account for differences in the hospital capture of features (Supplement: Table S1). Feature missingness of more than 10% resulted in the features being removed from the model. K-nearest neighbor (k = 5) imputation was used for missingness among the features with <10% missingness.[29] Additional information on the predictor transformations and feature missingness is included in the *Supplement: Missing Data.*

4. Facilities

We characterized the facilities included in the study to better understand the various contexts in which TL does and does not work for trauma outcomes modeling. Based on data available in the NTDB, facilities were described according to their size and the size of their trauma patient population. Proxies for resource availability included the number of ICU trauma beds and the number of trauma surgeons on staff. The patient cohort receiving trauma care at each facility was characterized by age, penetrating trauma prevalence, and injury severity.

Given the role of Level II trauma centers as tertiary centers for definitive trauma care,[30] we expected interfacility transfers to be prevalent within many facilities. Thus, the prevalence of primary admissions versus interfacility transfers within the trauma patient cohort was included as an important characteristic. Patients are transferred to higher levels of care due to resource needs or clinical acuity and may differ physiologically from primary admissions. To further assess the differences in patient encounters according to transfer status, we evaluated AIS sub scores (head, thorax, abdomen, and lower extremity) according to the prevalence of serious injury ($\geq$3) within each body region. The difference in the prevalence of serious injury was evaluated using the Chi^2 test and significance was defined if the primary admission cohort differed from the interfacility transfer cohort at the level of $p \leq .05$.

5. Statistical Analysis

Model Development

Logistic regression is a commonly used modeling approach in statistical analyses and machine learning for modeling a binary dependent variable and outputs the likelihood of the binary event. We developed a homogeneous TL model using logistic regression applied to the shared feature space using both source and target data. We optimized our models for the F1 score on a holdout set of target data. Given our interest in evaluating the model's performance on facility-specific target data, the holdout set contained only NTDB encounter data. Coefficients from multivariate analyses were estimated using logistic regression and were used to determine the strength of the predictor.[31] While a variety of ML algorithms could have been used in this evaluation, guidance in the literature recommends the use of less complex ML algorithms with TL techniques. For example, Curth *et al.* used logistic regression and gradient boosted classifiers to highlight the relative effect of TL on model performance.[22] Whereas other algorithms have numerous hyperparameters that can be tuned to improve model performance,

logistic regression does not. Therefore, the relative effect of TL on model performance may be more evident. In keeping with this guidance, we will only evaluate logistic regression-based models in this analysis.

As shown in Figure 3, the modeling was performed at the facility level. Patient records from each NTDB facility were combined with the eICU dataset to form the study cohort. Half of each facility's data were selected as the holdout set. The training set was then used to set the model parameters. We compared the distribution of eICU cohort characteristics to that of the NTDB cohort at the facility level across various features using the Mann-Whitney U test[32] and Fisher's exact test[33]. We evaluated the correlation between features for each facility. Features were considered aligned between facility specific records and eICU data if there was no significant difference (defined as p value >.05) between the distributions. Given the relatively small data size and to avoid overfitting, we constrained our feature sets by using LASSO.[34] LASSO, or Least Absolute Shrinkage and Selection Operator, is a regression-based approach to feature selection that aims to reduce redundancy across included features. LASSO does this by reducing the coefficients, often to zero, of features that do not contribute additional information to a predictive model. Similar to our approach with imputation, we executed our feature selection step during the facility-specific modeling process. While the potential features for inclusion were consistent across facilities, the final features included in the model differed according to their LASSO coefficients. After each model was evaluated and tuned on the validation set, the fit model was evaluated using the 2016 data from each facility, which served as the test set.

We bootstrapped our models and evaluated the average F1 score, 95% CI for recall, and 95% confidence interval for precision across 1000 model iterations. Bootstrapping in our modeling process involved sampling with replacement of the test set to better estimate the variability and performance of our models.[35] We used Python 3.8 for data management, model building, analysis, and evaluation.[36] Details of the Python packages used in these processes are available in the *Supplement: Detailed Methods.* R was used to map the ICD-9 diagnosis codes from the source data to Abbreviated Injury Severity scores.[26] ChatGPT was used throughout the exploratory data analysis and model development processes to assist with Python coding and code error mitigation.[37] No data were passed on to ChatGPT, nor did ChatGPT provide any analysis of the results.

6. *Model to Baseline Comparison*

To determine whether any benefit was conferred by TL, baseline models developed with target-only (NTDB) data were compared to those developed using TL methods. For each of the 30 selected facilities, we built three baseline models that differed in their included features (Table 3):

- the "all features" baseline: a target-only model composed of all available features in the target data,
- the "shared features" baseline: a target-only model composed of features in the shared feature space,
- the "NTS" baseline: a target-only model limited to features included in the New Trauma Score (NTS): age, systolic blood pressure, respiratory rate, GCS total, and ISS.[38]

The development of these models followed the same process as the homogeneous model development described above, with the following differences.

- The training and validation data consisted solely of target data. Twenty percent of each facility's training data served as the holdout set and to tune the LASSO penalty. The test data for all models consisted of facility-specific target data from 2016.
- The additional features included in the 'all features' baseline are listed in Table 3.
- LASSO regularization was not used for the NTS model due to the already limited feature set (n=5).

7. *Model Evaluation: Discrimination*

In each modeling stage, mortality during the inpatient encounter was defined as the positive class and labeled as f(x) = 1. Patient survival to discharge was defined as a negative class, labeled as f(x) = 0. The models output a probability for each patient encounter that was scored. Based on thresholds defined upon evaluation, the probability was translated into either 1 (for inpatient death) or 0 (for survival to discharge). The default threshold in scikit-learn packages is a predicted probability of >50% to be assigned a '1' and marked for the positive class. However, this method assumes a balanced dataset. In-hospital mortality after traumatic injury is highly imbalanced, with the prevalence of the positive class ranging between 5-10%. As such, we required a much lower threshold to define the positive class. During the modeling process, different thresholds were evaluated to determine the optimal threshold for each facility, with the range of possible values between 0.05 and 0.25.

After the encounters were scored, the predicted label was compared to the ground truth for each encounter. We evaluated all models using two types of metrics: discrimination and TL-specific metrics. Discrimination reflects how well an algorithm differentiates between two outcomes, such as sensitivity (recall), positive predictive value (precision), and F1 score, which is the harmonic mean of sensitivity and the positive predictive value.[39] The impact of TL at the facility level was assessed by comparing the best-scoring baseline model (via the F1 score) with the TL model. We defined the difference between these F1 scores, the TL Gap, and defined a Negative Transfer Gap as a negative score, indicating that the TL model performed worse than the baseline model.[17] We quantified the TL Gap in absolute terms (the difference between F1 scores) and relative terms (ratio of F1 scores). While generating these metrics, we focused on identifying the contexts in which TL models performed better than the target-only models. To do so, we explored trauma facility characteristics including interfacility transfer rates, patient population age, and severity of injured patients.

Results

1. *Participant Characteristics*

Source Data:

After removing 85 (1.2%) patient encounters for which no outcome could be ascertained, a total of 6,720 patient encounters composed the study's source data from the eICU, during which 499 patients died (7%). Among the source data encounters, the median age of surviving patients was 58 years, whereas the median age of patients who died was 71. Among surviving patients, 63% were male; 2/3 of patients who died were male. Fifty-nine percent of surviving patients required ventilator support, while 86% of those who died required ventilator support. GCS scores differed across groups, with a mean GCS total of 13.4 among survivors and 8.2 among patients who died. Comorbidities were common among both survivors and decedents (>50%) (Table 1).

Target Data:

Across the 30 included facilities, 16,140 patients received ICU care and met our inclusion criteria to compose the study cohort, of which 1518 died (8%). Among those classified as deceased in our evaluation, 199 were discharged for hospice care (13%). The median age of surviving patients was 51 years, whereas the median age of patients who died was 67. Among surviving patients, 67% were male; 65% of patients who died were male. Twenty-eight percent

of surviving patients required ventilator support, while 72% of those who died required ventilator support. GCS scores averaged 13.2 for the surviving cohort and 8.6 among those who died. Finally, 39% of the surviving patients reported comorbidities, compared to 55% of those who died (Table 1). Additional trauma-related characteristics are shown in Table 2, including a median ISS of 10 among survivors and 21 among patients who died.

Test Data:

NTDB data from 2016 served as the test data for our study. Across the 30 facilities in the study, 9,045 patient encounters were included, of which 921 (10.2%) died. Among surviving patients, the median age was 44 years, while the median age of patients that died was 46. Again, over 60% of surviving patients (66%) and deceased patients (63%) were male. 27% of surviving patients required a ventilator, while 75% of patients who died required ventilatory support. GCS scores averaged 12.9 for those who survived and 8.4 for those who died. Comorbidities were prevalent, with 38.6% of survivors and 53.7% of decedents reporting at least one comorbidity (Table 1). Table 2 also describes additional trauma-related characteristics of the test cohort, including that the median ISS of survivors was 12 and the median among those who died was 29.

The distribution of eICU cohort characteristics was compared with that of the NTDB cohort at the facility level across the following features (Table 4): age, sex, pulse, systolic blood pressure, ventilator requirement, GCS scores, and comorbidity. Alignment between a facility's patient characteristics and the eICU dataset was determined by the Mann-Whitney U test for continuous features and Fisher's exact test for binary features. Alignment for a feature was defined as a p-value greater than .05 on the test of interest, wherein a greater p-value indicates better alignment between the datasets. Sex was the most aligned feature, with 14/30 facilities having a distribution aligned with that of eICU cohort. Pulse and GCS (10/30 and 9/30, respectively) were the next most aligned features. Systolic blood pressure (2/30) was almost never aligned.

2. *Facility-Level Characteristics*

The 30 trauma facilities that were selected via stratified random sampling are listed in Table 5. Twelve facilities were "large" with $\geq$ 400 inpatient beds, and 18 "small" with <400 inpatient beds based on bed size groupings available in the target data. Large hospitals averaged 442 trauma admissions requiring ICU care per year, whereas small hospitals averaged 154. The mortality rate among these patients varied between facilities, with a minimum of 3.1% and a

maximum of 22.5%. The mean mortality rate among large facilities was 9.9% and the mean mortality rate among smaller facilities was 10.4%. The number of annual trauma deaths per facility ranged from 4 to 70. The median ISS scores ranged from 10-17 for large facilities and to 13-22 for smaller facilities. Age distributions varied across facilities, with some facilities having a younger patient population (median age 37, 39, 40 years) and some having an older population (median age 69, 65, and 63 years) (Supplemental Table S2).

The NTDB differentiates primary admissions from interfacility transfers. We described each NTDB facility according to the prevalence of interfacility transfer **to** that facility within the training data (Table 6). Interfacility transfer rates ranged from 0% to 61% with a median interfacility transfer rate of 19.5%. Interfacility transfer rates, on average, were similar across groups of small and large facilities with a mean of 23% and 24% respectively. In Table 6, we describe patient encounters within each facility according to their interfacility transfer status to evaluate differences in age, ISS, Abbreviated Injury Scores, and mortality rates. In over half of the facilities (n=17), age differed significantly between interfacility transfer patients and primary admissions. In 14 of these 17 facilities, the average age of the interfacility transferred patients was significantly greater than that of the primary admission patients. The prevalence of seriously injured body regions, defined as AIS$\geq$3, also differed by transfer status as shown in Table 6.

Within each facility, the distributions of characteristics in the NTDB training and testing cohort were compared within facilities using the Wald test. There was no significant difference ($p \leq .05$) in the distribution of data between each of the 30 facilities across the characteristics available.

3. Variable Selection

Across the shared feature space, and once combined with the eICU data, missing data were uncommon within each facility. Only once was a feature dropped from the modeling because of >10% missingness (oxygen saturation). Otherwise, all features were retained in the model, and imputation proceeded with the k-nearest neighbor imputation (k=5). Figures 4 and 5 illustrate the missingness by features across the source and target datasets, respectively. Supplemental Table S1 details feature missingness across the NTDB facilities.

Features were evaluated for correlation and features highly correlated with other features were excluded. GCS and its component scores were highly correlated with each other (>.8). The GCS motor score was retained, while GCS eye opening, GCS verbal, and GCS total scores were excluded for all facilities. We retained the GCS motor score, as it has been evaluated in

the literature as a significant predictor of mortality following trauma and linearly correlated with mortality, making it useful in regression modeling.[40] The binary indicator of advanced directives (DNR) was excluded from modeling for all facilities due to its nature as a proxy for the outcome of interest. We also evaluated the correlation between features and the outcome at the facility level. These correlation coefficients were compared to the correlation between features and the outcome in the eICU cohort (Supplemental Table S3). All features exhibited weak to very weak correlation to inpatient death following acute injury (correlation <.4) across nearly all facilities and the eICU cohort. LASSO regularization was used to reduce the dimensionality of the feature space and to reduce the likelihood of model overfitting (see *Supplement: Model Details*). Feature selection occurred at the facility level, so the selected features could adapt to different patterns and characteristics among the different facilities.

4. *Model Performance*

Overall, F1 scores were poor and ranged from 0.14 to 0.57. F1 scores are summarized here and in Table 7, with the full metrics available in Supplemental Table S4:

- Among facilities with a positive class less than 40 (n=14), TL scores average .31.
- Among facilities with a positive class larger than 40 (n=16), TL scores averaged .35.
- Correlation coefficients between F1 score and size of training cohort and size of positive class were weak (.173 and .164, respectively).
- Among facilities with less than 30% interfacility transfer rate (n=19), TL scores averaged .35, whereas among facilities with a 30% or higher interfacility transfer rate (n=11), TL scores averaged .29. The correlation coefficient between F1 score and patient transfer rate was moderately correlated at -0.44).

5. *Comparison to Baselines*

To evaluate the impact of TL, baseline models were constructed, and their performance metrics were compared with TL models. Each facility had three baseline models, as shown in Figure 3. F1 scores and comparisons to different models are shown in Table 7 and summarized here:

- The ‘all features’ baseline performed best overall, with an average F1 score of .34 across all 30 facilities, whereas the TL models averaged 0.33. The ‘all features’ baseline was the best performing model for 13 facilities.

- The TL model generated the highest F1 score for 12 facilities, with an average F1 score of 0.35. Compared to the 'shared features' baseline, the TL model performed better among 24 of 30 facilities (80%).
- Among facilities with an interfacility transfer rate <10% (n=6), TL always performed the best. Among facilities with an interfacility transfer rate ≥30% (n=11), TL F1 scores failed to surpass baseline models.

6. Post Hoc Analysis

After initial model results and comparisons to baseline proved underwhelming, we retrained our models and evaluated the impact of two alterations. First, we optimized our models for AUC instead of F1, to see if model performance varied when using a threshold-agnostic measure. To compare these results to baseline, we also evaluated those models when optimized for AUC. Second, we increased the relative influence of the target data within the training set by weighting it five times the base weight. Increasing the weight of the target data is a common approach in TL and Desautels et al. determined a weight of 5x to be most beneficial to their performance metrics.[17] Overall, AUC scores varied from 0.56 to 0.9 with a mean of 0.71. Compared to baselines, the homogeneous TL model performed relatively worse when using AUC to rank performance. The TL model outscored all models in 6 facilities (20%) and outscores the 'shared features' baseline in 9 facilities (30%). When evaluating models with weighted target data in the training set, F1 scores improved, on average, by 8%. However, some models suffered a reduction in performance when weighting the target data in the training set. When ranking modeling approaches compared to baselines, the TL model with weighted target data outperformed all baselines in 8 facilities (27%) and outperformed the 'shared features' baseline in 25 facilities (83%). These full metrics for these additional analyses are available upon request.

Discussion

We developed 30 facility-specific TL-based models to assess the likelihood of inpatient mortality after acute injury using a large and unlinked dataset to augment the number of training samples available for Level II trauma center modeling. Facility-specific modeling of a rare outcome leads to smaller record sets and an opportunity to examine the potential improvement in predictive performance using TL approaches. We compared the models developed with the TL approach to three baseline models. Across the facilities, the TL models exhibited poor

performance, with F1 scores ranging from 0.143 to 0.571, suggesting that mortality during trauma care is hard to predict.

The performances of the target-only baseline models were similarly poor and varied across a range like that of the TL models. Overall, the “all features” baseline model, which included features outside the shared feature space, performed better in more facilities than the TL model. The “all features” baseline consistently outperformed the other models which used subsets of the features included in the “all features” model. As expected, the models that included fewer features generally made less accurate predictions than those with a larger feature set. However, when comparing approaches limited to the shared feature space, the ‘shared features’ baseline and the TL approach, the TL approach tended to outperform the traditional ML approach (n=24).

We compared the performance metrics across all baseline models and evaluated the best-performing baseline for each facility to the TL model. We assessed the absolute and relative Transfer Learning Gaps and the TL approach performed better than or equal to the best baseline 40% of the time (n=12). When TL performed better, the average relative improvement in F1 scores was 33%. The relative performance of the TL model depended on the prevalence of interfacility transfers within a facility and absolute rarity, a function of the record set’s size and the outcome’s rarity. TL proved to be the best-performing model when interfacility transfer rates were low and when transfer rates were high, negative transfer occurred.

In our study, the prevalence of interfacility transfers proved to be an important characteristic determining the success of TL. The prevalence of interfacility transfer patients within a facility is likely a proxy for misalignment among patient characteristics and injury patterns. There are two proposed reasons for this. First, trauma patients that require transfer to a level II trauma center may be intrinsically different than trauma patients that are primary level II trauma patients. This difference may be due to age, type of injury, acuity of care required, among other factors. Second, the ‘time since injury’ varies among patients transferred to a trauma center. Consider the feature ‘systolic blood pressure – ED’. Among patients first treated at a level II trauma center, this vital sign record will consistently be an early data point in their care. Among patients transferred to a trauma center, this data point may be several hours further into the course of care. Effectively, transfer patients have an altered window of time post injury that is captured in the data compared to primary patients.

Our hypothesis that TL would perform best when record sizes were limited and outcomes were rare was inconsistently supported. A small record set was neither sufficient nor required for TL to improve performance relative to baseline models. Instead, additional

characteristics of the source and target data must be considered to determine the alignment of datasets. In this study, we attempted to confirm feature alignment through a homogeneous TL approach and feature engineering efforts. For example, the first recorded measurement of vitals and GCS scores in the eICU data was used to provide alignment with the first set of those features included in the NTDB. However, consider how the first capture of these features might differ for interfacility transfer patients. The first set of vitals and GCS scores captured in the NTDB is not the first set of hospital-collected measurements for interfacility transfer patients. Instead, a variable amount of time has transpired since they were first assessed at their hospital of origin. Indeed, in one study, the authors determined that the average time between interfacility trauma transfers was 186 minutes.[41] Therefore, the timing of these measurements among interfacility transfer patients may characterize a different stage of acute injury and the recovery course than in primary admissions. When time-varying features were examined for alignment between facilities and the eICU cohort, there was a slight difference based on transfer rates (Table 4). Among facilities with transfer rates <30% (n=19), alignment in at least one time-varying feature was seen in 74% of facilities. However, among facilities with transfer rates $\geq$30% (n=11), alignment in at least one time-varying feature was seen in 27% of facilities. Due to data constraints with the eICU dataset, we could not determine the prevalence of interfacility transfers within that data.

In addition to time-varying feature alignment across the source and target datasets, alignment of the feature space is another key consideration for TL. The “all features” baseline model and the NTS-based baseline model included ISS as a predictor, which was neither included in the “shared feature” baseline nor in the homogeneous TL model. ISS is an anatomical scoring system widely used to characterize patient injury; the score’s magnitude is associated with mortality following acute injury.[25,42] Other scoring systems exist to assess injury severity, including GCS, a physiologic scoring system, which was available and included in all models. While GCS scores are correlated with patient outcomes following injury, GCS scores lack the quantification of risk related to anatomical injury inherent in the ISS. Our work associated alignment between the eICU and facility-specific GCS scores with TL performance. Among the ten facilities where GCS scores were aligned with the eICU patient cohort, the average TL F1 score was .4 compared to a score of .3 among facilities where GCS scores were not aligned (n=20). The eICU dataset did not include ISS, nor did the included ICD 9 codes lend themselves to AIS scoring. Although some acuity of care characterization may have been provided by blood transfusion and ventilatory requirement features, the GCS score and its component scores were the only proxies for injury severity in the source dataset. Given the

strong correlation observed in the training data between GCS scores and inpatient death following acute injury, it makes sense that TL model performance was associated with GCS score alignment. Furthermore, incorporating other indicators for injury severity via more sophisticated TL methods may augment model performance in future iterations of this work.

One of the key considerations with TL is when to use it as a method over alternative ML methods.[15] In this work, we evaluated facility record sets that varied in numerous ways from the source data. Such irregularity included patient characteristics, the likelihood of inpatient mortality, the correlations between data set features and inpatient mortality, and characteristics specific to the facilities, like rates of interfacility transfers. Although the source data and the target data were alike in the overall characterization of patient care encounters following serious traumatic injury, this alone was inadequate to consistently improve predictive model performance with TL. Indeed, this lack of consistent association between source and target data elements was likely responsible for the overall poor performance of TL models. This finding is imperative and should be shared with personnel who may be faced with adopting or implementing TL methods at their institution.

Study Limitations

Our study had several limitations. First, our study relied on two retrospectively collected data sources designed for different purposes. The eICU database is archived medical data from the Philips ICU Telehealth platform and was developed to provide exploratory research related to ICU patient care.[23] In contrast, the NTDB was intended for trauma research.[24] Therefore, there are differences in data capture and integrity, which may have impacted our study results. For example, diagnosis coding is rigorously captured in the NTDB as it forms the basis for the ISS. Indeed, the NTDB initiated the National Trauma Data Standard in 2007, which has resulted in the NTDB being a highly standardized resource for trauma research.[24] eICU data were collected during the normal course of patient care across the represented hospitals. While the dataset has checksums for data integrity, the data have been minimally processed, leading to variability in completeness and reliability.[23] Such differences in the data collection and evaluation methods may have affected our research.

Additionally, the age of the data is a limitation of this work. Based on changes to NTDB data sharing, the ability to link facilities to each other in subsequent years was no longer available after 2016. This limited the recency of data we could use in this work. However, we do not expect the principles of TL or the conditions under which it is effective to be affected by data age.

Thirdly, the feature set used in the homogeneous TL model was constrained to the shared feature space and overall, the features had a weak correlation to the outcome of interest. Although the shared feature space defines our approach, we expected it to consist of features more strongly correlated with the outcome. We were disappointed with the sparsity of ICD codes in the eICU data, which prevented using injury descriptors in the model.

Finally, the alignment of the source and target datasets was a limitation of this study. When we evaluated the alignment between the source data and facility-specific target datasets across specific features, alignment was uncommon. The TL literature suggests methods to further align source and target datasets, such as reweighting source instances based on similarities to the target data.[16,17,43] There are many approaches to identifying similarity, including distance-based methods such as the Kullback-Leibler divergence or cosine similarity.

Conclusion

We experimented with an approach to augment Level II trauma center data with additional patient records from an unlinked dataset to improve prognostic modeling. We developed facility-specific predictive models using homogeneous TL methods. Although the performances of our models across all facilities were underwhelming, we were able to characterize trends and scenarios for when dataset augmentation with heterogenous data may be advantageous over more traditional ML approaches, such as when the size of the positive class is significantly constrained and interfacility transfer rates are low. Our conclusions regarding the types of facilities that experienced model improvement will be useful in future work, including evaluating the impact of heterogeneous TL in this domain. The potential impact of this study illustrates a critical evaluation of TL and assesses these advanced modeling techniques among facilities with constrained record sets.

References

1. Raghupathi W, Raghupathi V. Big Data Analytics in healthcare: Promise and potential. *Health Information Science and Systems*. 2014;2(1). doi:10.1186/2047-2501-2-3

2. Jordan MI, Mitchell TM. Machine learning: Trends, Perspectives, and prospects. *Science*. 2015;349(6245):255-260. doi:10.1126/science.aaa8415

3. Yang J, Soltan AA, Clifton DA. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *npj Digital Medicine*. 2022;5(1). doi:10.1038/s41746-022-00614-9

4. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*. Published online 2021. doi:10.1001/jamainternmed.2021.2626

5. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in Clinical Research and machine learning in Health Care. *The Lancet Digital Health*. 2020;2(9). doi:10.1016/s2589-7500(20)30186-2

6. Krumholz HM. Big Data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*. 2014;33(7):1163-1170. doi:10.1377/hlthaff.2014.0053

7. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples. *ACM Computing Surveys*. 2020;53(3):1-34. doi:10.1145/3386252

8. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: A synthetic knowledge synthesis. *Science Progress*. 2022;105(1):003685042110297. doi:10.1177/00368504211029777

9. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLOS ONE*. 2019;14(11). doi:10.1371/journal.pone.0224365

10. Al-Stouhi S, Reddy CK. Transfer Learning for Rare Class Analysis.

11. Haibo He, Garcia EA. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. 2009;21(9):1263-1284. doi:10.1109/tkde.2008.239

12. El-Banna M. Modified mahalanobis taguchi system for imbalance data classification. *Computational Intelligence and Neuroscience*. 2017;2017:1-15. doi:10.1155/2017/5874896

13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357. doi:10.1613/jair.953

14. Al-Stouhi S, Reddy CK. Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems*. 2015;48(1):201-228. doi:10.1007/s10115-015-0870-3

15. Pan SJ, Yang Q. A survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2010;22(10):1345-1359. doi:10.1109/tkde.2009.191

16. Weiss K, Khoshgoftaar TM, Wang D. A survey of Transfer Learning. *Journal of Big Data*. 2016;3(1). doi:10.1186/s40537-016-0043-6

17. Wang Z, Dai Z, Poczos B, Carbonell J. Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2019. doi:10.1109/cvpr.2019.01155

18. Singh L, Janghel RR, Sahu SP. A boosting-based transfer learning method to address absolute-rarity in skin lesion datasets and prevent weight-drift for melanoma detection. *Data Technologies and Applications*. 2022;57(1):1-17. doi:10.1108/dta-10-2021-0296

19. Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. *2012 IEEE 12th International Conference on Data Mining Workshops*. Published online 2012. doi:10.1109/icdmw.2012.93

20. Wiens J, Guttag J, Horvitz E. A study in Transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*. 2014;21(4):699-706. doi:10.1136/amiajnl-2013-002162

21. Desautels T, Calvert J, Hoffman J, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical Informatics Insights*. 2017;9:117822261771299. doi:10.1177/1178222617712994

22. Curth A, Thoral P, van den Wildenberg W, et al. Transferring clinical prediction models across hospitals and Electronic Health Record Systems. *Machine Learning and Knowledge Discovery in Databases*. Published online 2020:605-621. doi:10.1007/978-3-030-43823-4_48

23. 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). doi:10.1038/sdata.2018.178

24. Hashmi ZG, Kaji AH, Nathens AB. Practical guide to surgical data sets: National Trauma Data Bank (NTDB). *JAMA Surgery*. 2018;153(9):852. doi:10.1001/jamasurg.2018.0483

25. Baker SP, O'Neill B, Haddon W, Long WB. The injury severity score. *The Journal of Trauma: Injury, Infection, and Critical Care*. 1974;14(3):187-196. doi:10.1097/00005373-197403000-00001

26. Clark DE, Black AW, Skavdahl DH, Hallagan LD. Open-access programs for injury categorization using ICD-9 or ICD-10. *Injury Epidemiology*. 2018;5(1). doi:10.1186/s40621-018-0149-8

27. Kozar RA, Holcomb JB, Xiong W, Nathens AB. Are all deaths recorded equally? the impact of hospice care on risk-adjusted mortality. *Journal of Trauma and Acute Care Surgery*. 2014;76(3):634-641. doi:10.1097/ta.0000000000000130

28. American College of Surgeons. National Trauma Data Bank National Trauma Data Standard. Data Dictionary. 2016 Admissions. Available at: https://www.facs.org/~/media/files/quality%20programs/trauma/ntdb/ntds/data%20dictionaries/ntds%20data%20dictionary%202016.ashx

29. Batista GE, Monard MC. A study of K-nearest neighbour as an imputation method. *His.* 2002;87(48):251-260.

30. American College of Surgeons Committee on Trauma. Resources for the Optimal Care of the Injured Patient. Chicago, IL: American College of Surgeons, 2014.

31. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: An overview. *Journal of Thoracic Disease*. 2019;11(S4). doi:10.21037/jtd.2019.01.25

32. McKnight PE, Najab J. Mann-Whitney U Test. *The Corsini Encyclopedia of Psychology*. Published online 2010:1-1. doi:10.1002/9780470479216.corpsy0524

33. Kim H-Y. Statistical notes for clinical researchers: Chi-Squared Test and Fisher's exact test. *Restorative Dentistry & Endodontics*. 2017;42(2):152. doi:10.5395/rde.2017.42.2.152

34. Muthukrishnan R, Rohini R. Lasso: A feature selection technique in predictive modeling for Machine Learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. Published online 2016. doi:10.1109/icaca.2016.7887916

35. Nakatsu RT. An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems*. 2021;36(3):51-57. doi:10.1109/mis.2020.2978066

36. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*. 2011;12:2825-30.

37. OpenAI. ChatGPT, 2023. [Large Language Model]

38. Jeong JH, Park YJ, Kim DH, et al. The New Trauma Score (NTS): A modification of the revised trauma score for better trauma mortality prediction. *BMC Surgery*. 2017;17(1). doi:10.1186/s12893-017-0272-4

39. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models. *JAMA*. 2017;318(14):1377. doi:10.1001/jama.2017.12126

40. Healey C, Osler TM, Rogers FB, et al. Improving the Glasgow Coma Scale Score: Motor Score Alone is a better predictor. *The Journal of Trauma: Injury, Infection, and Critical Care*. 2003;54(4):671-680. doi:10.1097/01.ta.0000058130.30490.5d

41. Garwe T, Cowan LD, Neas B, Cathey T, Danford BC, Greenawalt P. Survival benefit of transfer to tertiary trauma centers for major trauma patients initially presenting to Nontertiary Trauma Centers. *Academic Emergency Medicine*. 2010;17(11):1223-1232. doi:10.1111/j.1553-2712.2010.00918.x

42. Orhon R, Eren SH, Karadayi S, et al. Comparison of trauma scores for predicting mortality and morbidity on trauma patients. *Turkish Journal of Trauma and Emergency Surgery*. 2014;20(4):258-264. doi:10.5505/tjtes.2014.22725

43. Daumé III H. Frustratingly easy domain adaptation. *arXiv.* Preprint posted online July 10, 2009. arXiv:0907.1815.

Tables

Table 1. eICU and NTDB Training Cohort Descriptions (2014, 2015).

	eICU				NTDB							
	Study Cohort (2014, 2015)								Test Cohort (2016)			
	Survived (n=6221)		Died (n=499)		Survived (n=14622)		Died (n=1518)		Survived (n=8124)		Died (n=921)	
Feature	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Age	58	22.5	71	23.8	51	22.9	67	22.7	44	37.1	46	43.5
Sex	.63	.483	.66	.473	.67	.469	65%	.476	.66	.475	.63	.483
Syst BP	128	25	123	31	137	28.5	138	41.9	136	33.6	138	47.9
Pulse	85	19	87	24	88	22.4	88	28.8	88	24.8	86	31.65
Oxy Sat	97	3.7	96	7.4	97	4.6	95	9.1	93.8	17.4	88.6	25.5
Vent Bin	.595	.491	.861	.346	.276	.447	.724	.447	.27	.444	.746	.435
GCS eye	3.59	.9	2.26	1.4	3.6	.99	2.4	1.4	3.44	1.31	2.3	1.67
GCS verb	4.2	1.4	2.38	1.7	4.22	1.34	2.6	1.8	4.08	1.66	2.52	2.01
GCS motor	5.6	1.2	3.52	2.2	5.4	1.45	3.5	2.3	5.24	1.84	3.39	2.53
GSC total	13.39	3.18	8.17	5.04	13.2	3.6	8.6	5.4	12.93	4.14	8.38	5.7
Blood trans	.024	.152	4.6	.21	.047	.211	.105	.306	.007	.085	.014	.118
Comorbid	.537	.5	.571	.5	.392	.488	.549	.498	.386	.487	.537	.499
Hypertension	.33	.47	.37	.48	.31	.46	.43	.495	.312	.464	.436	.496
Diabetes	.144	.35	.15	.35	.12	.32	.16	.364	.132	.338	.15	.357
CHF	.058	.23	.094	.292	.03	.17	.07	.259	.032	.176	.089	.285
CVA	.057	.23	.068	.252	.02	.15	.05	.225	.026	.158	.064	.245
COPD	0	0	0	0	0	0	0	0	.058	.234	.1	.3
MI	.043	.20	.056	.23	.01	.11	.018	.135	.009	.094	.014	.118
Renal	.041	.197	.068	.252	.01	.1	.028	.166	.016	.126	.051	.22
Cancer	.06	.24	.08	.27	.007	.08	.020	.141	.009	.094	.02	.139
Internal organ	.577	.49	.709	.45	.70	.458	.713	.452	.736	.441	.759	.428
Open wounds	.08	.28	.07	.26	.367	.482	.307	.461	.356	.479	.295	.456
Fracture	.47	.5	.44	.5	.610	.488	.605	.49	.628	.483	.631	.482
Blood vessels	.02	.14	.016	.126	.044	.205	.06	.237	.05	.217	.05	.218
Sprains	0	.03	0	0	.059	.236	.02	.139	.067	.249	.028	.166
Dislocation	.0106	.10	.004	.063	.048	.214	.035	.184	.063	.243	.056	.231
Amputation	0	0	0	0	.014	.119	.007	.085	.004	.062	.011	.104
Other	.065	.25	.144	.35	0	0	0	0	.027	.161	.017	.131

Table 2. Description of Additional NTDB Characteristics

	Study Cohort (2014, 2015)				Test Cohort (2016)			
	Survived (n=14622)		Died (n=1518)		Survived(n=8124)		Died (n=921)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ISS	10	10	21	14.4	12	9.27	29	13.86
Type of Injury								
Blunt	.863	.344	.855	.352	.879	.327	.866	.34
Penetrating	.063	.243	.072	.258	.066	.248	.064	.245
Burn	.001	.038	.003	.057	.001	.037	.001	.033
Other	.031	.174	.036	.185	.014	.117	.011	.104
Mechanism of Injury								
Fall	.403	.49	.552	.497	.412	.492	.561	.496
MVC	.326	.469	.25	.433	.331	.471	.239	.427
Firearm	.033	.178	.063	.242	.036	.186	.055	.229
Struck	.059	.235	.022	.146	.087	.282	.023	.149
Machinery	.003	.057	.003	.051	.002	.041	0	0
Transport	.072	.259	.028	.051	.058	.234	.033	.18
Other	.028	.165	.038	.192	.04	.196	.058	.233
Intent of Injury								
Unintentional	.843	.364	.856	.351	.847	.360	.86	.347
Assault	.089	.285	.047	.211	.089	.285	.033	.18
Self-Inflicted	.019	.137	.053	.225	.017	.13	.04	.196
Undetermined	.005	.07	.006	.077	.007	.083	.01	.098
Other	.003	.054	.003	.057	.027	.161	.017	.131

Table 3. Feature Lists

Feature	TL Shared	BL All	BL Shared	BL NTS
Age	✓	✓	✓	✓
Sex	✓	✓	✓	
BMI	✓	✓	✓	
Systolic BP	✓	✓	✓	✓
Pulse	✓	✓	✓	
Temp		✓		
Oxygen Saturation	✓	✓	✓	
Supplemental Oxygen			✓	
Respiratory Rate		✓		✓
Ventilator Use	✓	✓	✓	
GCS-eye	✓	✓	✓	
GCS-verb	✓	✓	✓	
GCS-motor	✓	✓	✓	
GCS-total	✓	✓	✓	✓
ISS			✓	✓
Blood-trans	✓	✓	✓	
Type of transport to ED		✓		
Comorbidity (binary)	✓	✓	✓	
Hypertension	✓	✓	✓	
Diabetes	✓	✓	✓	
CHF	✓	✓	✓	
CVA	✓	✓	✓	
COPD	✓	✓	✓	
MI	✓	✓	✓	
Renal Disease	✓	✓	✓	
Cancer	✓	✓	✓	
Nature of Injury	✓	✓	✓	
Mechanism of Injury		✓		
Intent of Injury		✓		
Type of Injury		✓		

Table 4. Evaluation of Alignment Between Characteristics of eICU Encounters to NTDB by Facility`

Facility	Age#	Sex&	Pulse#	Syst BP#	Vent&	GCS-#	Comorbid&
6547tl	*	.61	.81	*	*	.57	*
2194tl	.7	.16	.48	*	*	.2	*
355tl	*	.7	*	*	*	.01	*
6268tl	*	.67	.13	*	*	.03	*
6552tl	.07	.59	.02	*	*	.22	.07
6555tl	*	.01	.04	*	*	.43	*
2005tl	*	*	*	*	*	.05	*
2681tl	.1	.46	.26	*	*	.01	.02
2705tl	.2	.39	.43	*	*	*	.58
2416tl	*	.19	*	*	*	*	*
6242tl	*	*	*	*	*	*	*
2399tl	.02	.45	.16	*	*	.09	.04
2380sh	*	*	*	*	*	*	.08
6362all	.01	.05	.03	*	*	.23	*
138all	*	*	*	*	*	.18	*
2254all	*	.01	.01	*	*	.79	*
2310all	*	*	*	*	*	.01	*
5003NTS	*	*	*	.3	.85	*	*
2404NTS	*	.03	*	*	*	*	.01
2065NTS	*	.37	*	.8	*	*	*
2154all	.06	.04	.39	*	*	*	.01
6244all	*	*	*	*	*	*	*
358all	*	*	.97	*	*	*	*
5004all	*	*	*	*	.61	*	*
2001all	.5	.61	.11	*	*	.1	.11
2199all	*	.15	*	*	*	*	*
460all	*	*	*	*	*	*	*
315NTS	.12	.27	.1	*	*	.04	.43
2680all	*	.07	*	.03	*	.03	*
6131all	*	*	*	*	.02	*	.01

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best; *all* denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best

# denotes significance testing using Mann Whitney U test; & denotes significance testing using Fischer's Exact test

*denotes a p-value < .01

Table 5. NTDB Facility Descriptions, 2014-2015

Facility	Cohort Size	Positive Class	Transfer Rate (%)	ICU Trauma Beds	Trauma Surgeons	Region	Teaching Status
6547[tl]	435	38	11.3	16-25	5-6	South	Non-teaching
2194[tl]	126	14	0	11-15	4-5	Northeast	Community
355[tl]	1037	82	6.8	>35	7-8	West	Community
6268[tl]	214	20	15.4	16-25	4-5	West	Non-teaching
6552[tl]	401	22	0.7	26-35	4-5	Northeast	Community
6555[tl]	232	36	0.9	16-25	4-5	Midwest	Community
2005[tl]	787	74	9.4	16-25	4-5	West	Community
2681[tl]	463	37	16.6	16-25	5-6	West	Community
2705[tl]	212	8	5.7	16-25	4-5	South	Community
2416[tl]	580	57	26.6	11-15	7-8	West	Community
6242[tl]	424	13	22.4	26-35	5-6	West	Community
2399[tl]	660	92	24.1	>35	5-6	Northeast	University
2380[sh]	792	75	33.5	>35	5-6	Northeast	University
6362[all]	471	42	15.3	16-25	4-5	Midwest	Community
138[all]	1097	76	13.9	>35	4-5	West	Community
2254[all]	780	61	15.9	16-25	4-5	West	Non-teaching
2310[all]	421	23	12.4	11-15	4-5	West	Community
5003[NTS]	119	15	40.3	11-15	4-5	Midwest	Non-teaching
2404[NTS]	436	61	29.8	26-35	5-6	Midwest	Non-teaching
2065[NTS]	468	63	61.3	>35	>8	Midwest	Community
2154[all]	269	32	32.7	26-35	7-8	Midwest	Non-teaching
6244[all]	1316	137	15.1	16-25	5-6	South	Community
358[all]	1439	69	9.85	16-25	4-5	West	Non-teaching
5004[all]	147	13	49.7	11-15	>8	Midwest	University
2001[all]	351	44	41.3	11-15	4-5	Midwest	Community
2199[all]	678	140	43.4	26-35	4-5	Midwest	Non-teaching
460[all]	868	59	47.4	>35	4-5	West	Community
315[NTS]	204	37	21.6	26-35	4-5	Midwest	Non-teaching
2680[all]	575	60	33.9	1-10	1-3	Midwest	Community
6131[all]	138	31	47.1	11-15	4-5	West	Community

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best; *all* denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best

Table 6. Differences in Transfer vs Primary Admission Patients by Facility, 2014-2015

Facility	Transfer Rate (%)	Patient Encounters		Patient Age#		Mortality (%)&		ISS#		AIS-head ≥3(%)&		AIS-thorax ≥3(%)&		AIS-abdomen ≥3(%)&		AIS-LE ≥3 (%)&	
		Primary	Transfer	Primary	Transfer	Primary	Transfer	Primary	Transfer	Primary	Transfer	Primary	Transfer	Primary	Transfer	Primary	Transfer
6547[tl]	11.3	386	49	51.2#	59.7	9.3	4.1	18.7	16.5	0.57	0.59	0.22	0.22	0.12	0.10	0.15	0.08
2194[tl]	0	126	0	57.8	n/a	11.1	n/a	16.3	na	0.74	n/a	0.13	n/a	0.04	n/a	0.06	n/a
355[tl]	6.8	966	71	49.5	48.0	8.4	1.4	14.9	13.2	0.37	0.37	0.14	0.10	0.04	0.07	0.08&	0.00
6268[tl]	15.4	181	33	53.5	50.3	8.3	15.2	17.1	19.8	0.59	0.48	0.28	0.39	0.06	0.12	0.12	0.06
6552[tl]	0.7	398	3	54.9	53.0	5.5	0	16.5	22.3	0.53	0.67	0.25	0.00	0.11	0.00	0.07	0.33
6555[tl]	0.9	230	2	64.0	73.5	15.7	0	16.2	19.5	0.55	1.00	0.16	0.50	0.03	0.00	0.11	0.00
2005[tl]	9.4	713	74	46.4	49.1	9.4	9.5	15.8	15.4	0.37	0.35	0.22	0.23	0.06	0.08	0.10	0.07
2681[tl]	16.6	386	77	58.7	61.8	8.8	3.9	16.9	17.6	0.53&	0.69	0.26&	0.12	0.05	0.05	0.09	0.06
2705[tl]	5.7	200	12	58.9	64.3	4	0	13.5	16	0.43	0.58	0.27	0.42	0.04	0.00	0.15	0.00
2416[tl]	26.6	426	154	51.9#	57.1	10.1	9.1	15.8	14.3	0.44	0.49	0.28&	0.19	0.09	0.05	0.12	0.13
6242[tl]	22.4	329	95	39.7#	48.4	3	3.2	20.9#	16.6	0.39	0.38	0.50&	0.33	0.22&	0.11	0.15&	0.05
2399[tl]	24.1	501	159	57.3#	66.2	14.4	12.6	18.4	17.0	0.60&	0.75	0.20&	0.11	0.07&	0.02	0.10	0.05
2380[sh]	33.5	527	265	48.0#	56.5	9.7	9.1	19.8#	15.6	0.54	0.56	0.30&	0.18	0.15	0.13	0.10&	0.03
6362[all]	15.3	399	72	53.4#	59.4	9	8.3	17.9	15.7	0.56&	0.72	0.21&	0.03	0.08	0.03	0.14&	0.03
138[all]	13.9	944	153	49.6#	58.6	7	6.5	17.5#	15.6	0.50	0.53	0.23	0.17	0.10	0.09	0.05	0.05
2254[all]	15.9	656	124	52.3#	57.3	8.2	5.6	17.4	15.1	0.52&	0.67	0.24&	0.06	0.08	0.10	0.11	0.06
2310[all]	12.4	369	52	43.3	39.5	4.6	11.5	14.4	14	0.31	0.29	0.28	0.25	0.09	0.13	0.16	0.10
5003[NTS]	40.3	71	48	39.8	39.3	12.7	12.5	16.8	20.1	0.38	0.46	0.28	0.35	0.14	0.10	0.14	0.17
2404[NTS]	29.8	306	130	49.1	52.3	13.7	14.6	16.3	16.9	0.46&	0.64	0.31&	0.15	0.11&	0.02	0.09&	0.03
2065[NTS]	61.3	181	287	48.8	46.2	16	11.8	18.1	17.8	0.52	0.53	0.29	0.29	0.09	0.10	0.16	0.11
2154[all]	32.7	181	88	53.1	57.5	11	13.6	20.1#	16.8	0.50	0.42	0.30	0.35	0.14	0.11	0.10	0.07
6244[all]	15.1	1117	199	47.6#	58.4	10.7	8.5	15.5	16	0.40&	0.62	0.23&	0.11	0.08	0.07	0.13&	0.07
358[all]	9.85	1298	141	49.9#	55.6	5	2.8	16.7	16.3	0.33&	0.42	0.36&	0.26	0.08	0.12	0.09	0.07
5004[all]	49.7	74	73	49.4#	43.2	10.8	6.8	21	20.1	0.46	0.58	0.36	0.30	0.15	0.12	0.11	0.08
2001[all]	41.3	206	145	55.6#	61.6	12.1	13.1	17	15.6	0.44&	0.58	0.26&	0.12	0.06	0.06	0.17&	0.06
2199[all]	43.4	384	294	45.6#	50.5	20.8	20.4	18.7	16.9	0.40&	0.56	0.32&	0.20	0.16&	0.09	0.16&	0.06
460[all]	47.4	457	411	44.2#	47.9	7.2	6.3	17.3#	15.8	0.45&	0.58	0.33	0.17	0.10&	0.04	0.09	0.07
315[NTS]	21.6	160	44	52.2#	63.7	11.9	11.4	13.1	122	0.31	0.39	0.22	0.14	0.07	0.09	0.18	0.20
2680[all]	33.9	80	195	53.8#	44.2	10.8	9.7	11.5	12.4	0.34	0.43	0.19	0.24	0.06	0.10	0.17&	0.07
6131[all]	47.1	73	65	47.7#	37.9	32.9&	10.8	22.6	23.8	0.53	0.62	0.47	0.48	0.11	0.14	0.11	0.09

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best.

all denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best.

Independent t-tests used to compare distributions between primary admissions and interfacility transfers; # signifies a p-value ≤ .05.

& Chi^2 tests used to compare distributions between primary admissions and interfacility transfers; & signifies a p-value ≤ .05.

Table 7. F1 Scores and Transfer Learning Gaps across Modeling Approaches by Facility

Facility	BL-all	BL-shared	BL-NTS	TL	Best Model	Absolute Learning Gap	Relative Learning Gap
6547	0.207	0.194	0.222	0.379	TL	0.157	71%
2194	0.31	0.31	0.348	0.571	TL	0.223	64%
355	0.194	0.226	0.204	0.353	TL	0.127	56%
6268	0.143	0.136	0.3	0.414	TL	0.114	38%
6552	0.32	0.309	0.169	.436	TL	0.116	36%
6555	0.22	0.222	0.22	0.295	TL	0.073	33%
2005	0.359	0.380	0.291	0.5	TL	0.12	32%
2681	0.191	0.167	0.15	0.246	TL	0.055	29%
2705	0.168	0.165	0	0.207	TL	0.039	23%
2416	0.192	0.243	0.255	0.286	TL	0.031	12%
6242	0.209	0.118	0.113	0.216	TL	0.007	3%
2399	0.333	0.338	0.331	0.339	TL	0.001	0.30%
2380	0.376	0.379	0.141	0.362	BL-Shared	-0.017	-4%
6362	0.435	0.332	0.35	0.412	BL-All	-0.023	-5%
138	0.392	0.27	0.305	0.370	BL-All	-0.022	-6%
2254	0.429	0.221	0.356	0.386	BL-All	-0.043	-10%
2310	0.357	0.214	0.262	0.316	BL-All	-0.041	-11%
5003	0.118	0.105	0.175	0.143	BL-NTS	-0.032	-18%
2404	0.443	0.459	0.475	0.389	BL-NTS	-0.086	-18%
2065	0.26	0.258	0.374	0.301	BL-NTS	-0.073	-20%
2154	0.410	0.309	0.273	0.321	BL-All	-0.089	-22%
6244	0.444	0.327	0.368	0.339	BL-All	-0.105	-24%
358	.468	.411	.301	.333	BL-All	-0.135	-29%
5004	0.5	0.28	0.281	0.353	BL-All	-0.147	-29%
2001	0.453	0.281	0.4	0.318	BL-All	-0.135	-30%
2199	0.502	0.334	0.35	0.346	BL-All	-0.156	-31%
460	0.382	0.231	0.4	0.274	BL-All	-0.126	-32%
315	0.491	0.490	0.467	0.31	BL-NTS	-0.181	-37%
2680	0.434	0.287	0.289	0.275	BL-All	-0.159	-37%
6131	0.316	0.2	0.4	0.143	BL-All	-0.257	-64%
mean	0.34	0.27	0.29	0.33			

Figure Legends

Figure 1. Flowchart of the eICU study population. From the entirety of the patient unit stay IDs in the dataset, the cohort was filtered to those with 'trauma' in the diagnosis string. Encounters with non-acute trauma diagnoses were also excluded. The results were flattened to hospital stay. Hospital stays with missing outcome data were removed from the data set.

Figure 2. Flowchart of the NTDB Study Population. From the entire NTDB database of the 2014 and 2015 encounters, 30 Level II facilities were chosen via stratified random sampling based on hospital size. The cohort was determined by patients who required critical care during the first 24 hours of their hospital stay. Modeling and evaluation were performed at the facility level. Facility-specific NTDB data split with 50% of the data were added to the source data to form the training set, and 50% of the data were used for the validation set.

Figure 3. Modeling Schematic of Three Specific Aims. The first aim of this study was to develop and validate facility-specific homogeneous transfer-learning models to predict inpatient mortality following traumatic injury. Each facility-specific model includes source and facility-specific target data in the training set and only facility-specific target data in the validation set. Facility-specific data from 2016 was used to compose the test set. The feature space is defined by the shared feature space. The second aim involves the development and validation of a heterogeneous transfer learning approach, which is discussed in the next section of this work. Aim 3 involved the development and testing of three facility-specific baseline models for comparison with the transfer learning models in Aim 1. The baseline models will be composed of only facility-specific data from the target source for training, validation, and testing.

Figure 4. Percentage of Missingness among Features (eICU): All features used in Aim 1 are listed according to their percentage of missingness. All features had less than 5% missing data.

Figures

Figure 1. Flowchart of eICU Study Population

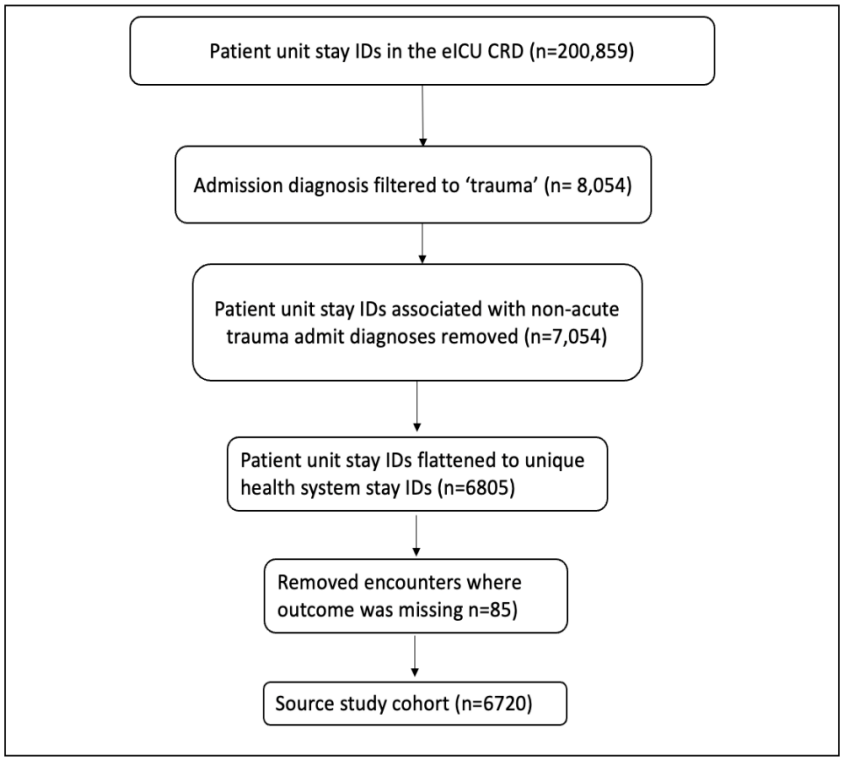


Figure 2. Flowchart of NTDB Study Population

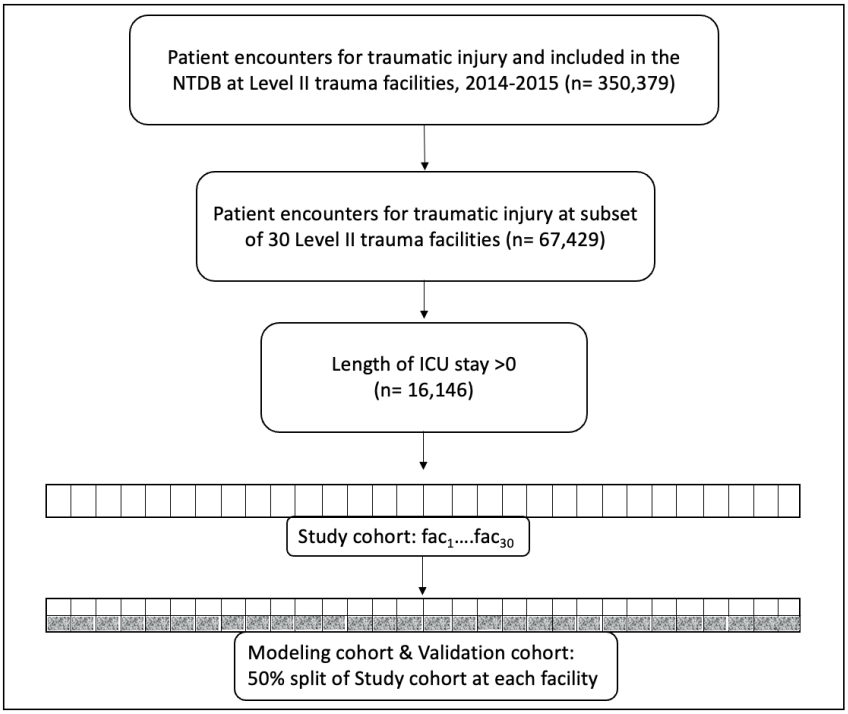


Figure 3. Modeling Schematic of TL models compared to baseline models

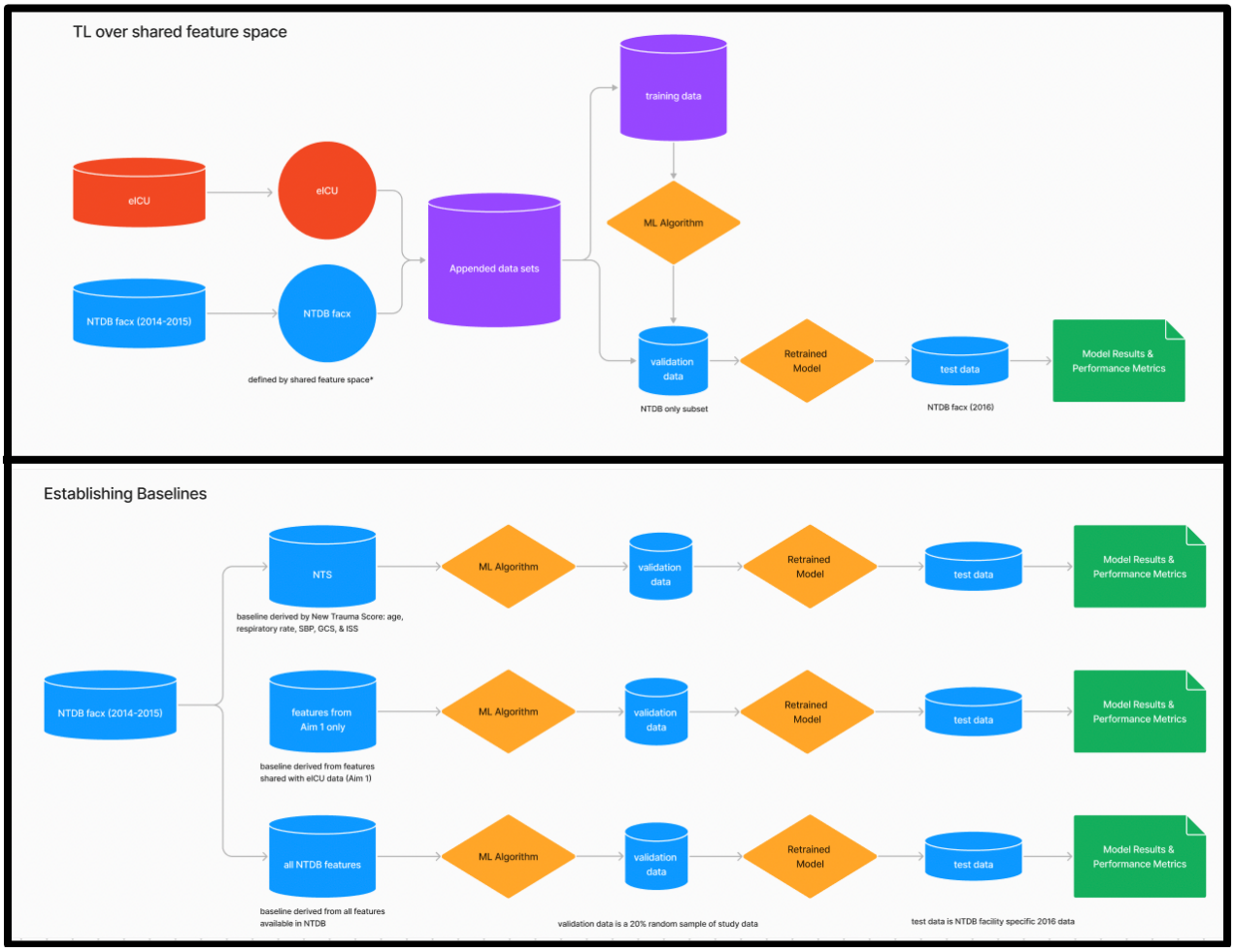
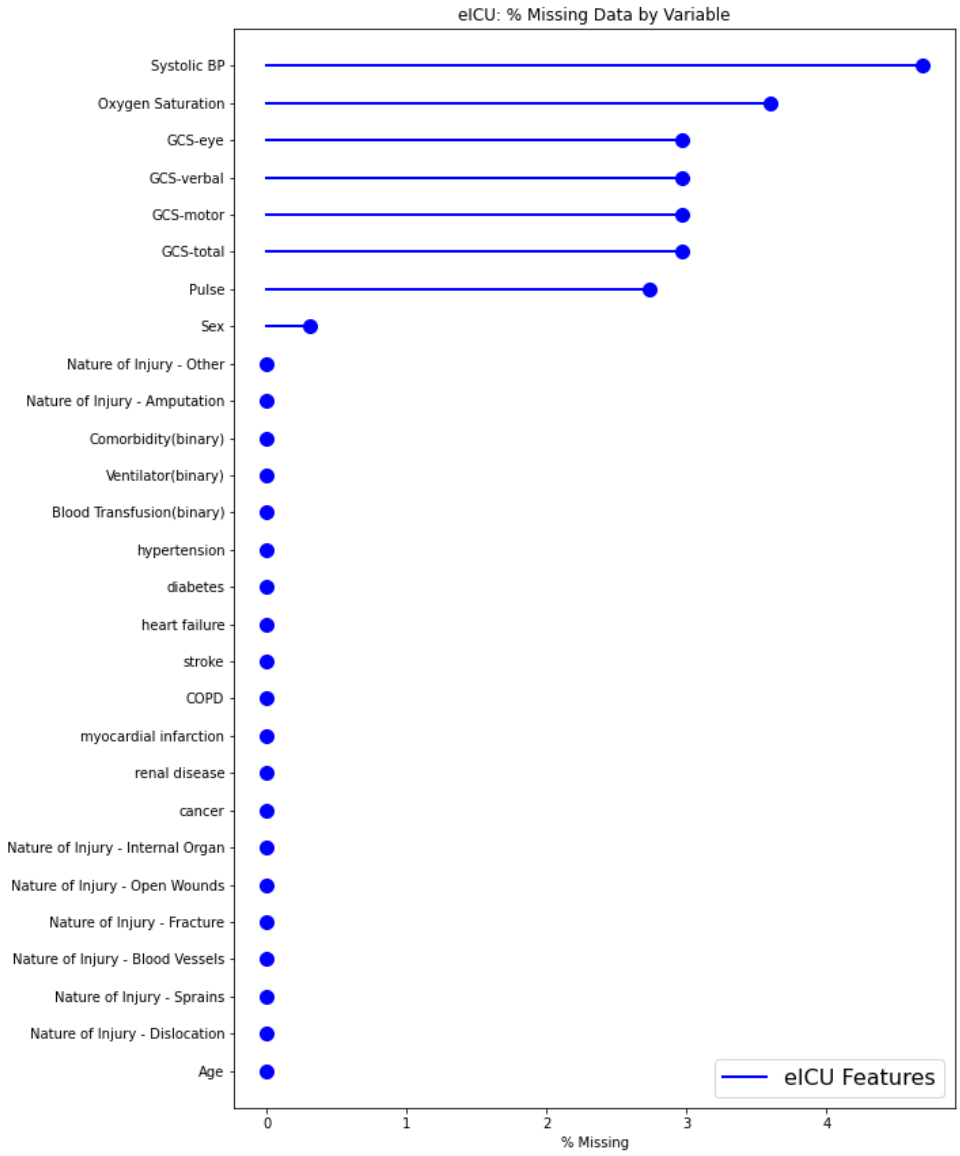


Figure 4. Proportion of Missingness Among Features, eICU 2014-2015



Supplemental Material

Supplemental Paper 1

eICU Cohort Determination

As described, the cohort of trauma encounters was determined based on the string 'trauma' in the diagnosis field of the eICU data. However, when evaluating the entirety of the diagnosis string, it became evident that non-traumatic diagnoses remained in the data. The following diagnoses and their associated encounters were excluded from the study cohort:

- Non-traumatic amputation (n=126)
- Facial surgery (n=59)
- Non-traumatic fracture (n=124)
- Non-traumatic hemorrhage (n=338)
- Hip surgery (n=126)
- knee surgery (n=83)
- non-traumatic coma (n=11)
- trauma medical condition (n=83)

Feature Development

In homogeneous TL, only the features shared across the source and target datasets were used. We engineered the features to increase the number of shared features available for homogeneous TL. The following list describes each feature and the calculations or groupings required for the source and target data to align them in the shared feature space:

- Outcome:
    - eICU: discharge location of death served as outcome = 1
    - NTDB: discharge disposition of death or hospice served as outcome = 1
- Age: maintained as a continuous variable
    - eICU: In the original data, patients older than 89 years are indicated as age >89 years. This was modified to age = 90 in the dataset
    - NTDB: patient ages in months were all rounded to "0" or "1." Patients aged "≥ 89 years were modified to 90 years.
- Sex
    - eICU: binary variable including male (1) or female (2)
    - NTDB: binary variable including male (1) or female (2)
- Pulse: maintained as a continuous variable
- Systolic Blood Pressure (SBP): maintained as a continuous variable
- Oxygen Saturation (oxysat): maintained as a continuous variable with a maximum value of 100
    - eICU: component scores and composite obtained from (table / field)
    - NTDB: component scores and composite obtained from (table/ field)
- Body mass index (BMI): once calculated, maintained as a continuous variable
    - eICU: calculated as weight in kg divided by the square of height in meters; outliers were discovered based on these calculations and treated as follows:
        - values less than x:
        - values more than 100: treated as missing (see below)
        - values > 60 and <100, respectively, and modified to 60.
    - NTDB: calculated as weight in kg divided by square of height in meters
        - values less than x:
        - values more than 100: treated as missing (see below)
        - values > 60 and <100, respectively, and modified to 60.
- Glasgow Coma Scale (GCS): treated as an integer
    - eICU: component scores and composite obtained from (table / field)
    - NTDB: component scores and composite obtained from (table/ field)
- blood transfusion: binary indicator defined by transfusion of red blood cells within 24 hours of admission
    - eICU: transfusion defined as (what table / field) with text string including "packed red blood cells", "transfusion of >2 units prbcs", "blood", or "transfusion of 1-2 units of prbcs" (within the prescribed offset of 1440)
    - NTDB: transfusion defined within table 'PCODEDESCR' as "transfusion of paced cells" on day 0 or 1 of admission.
- Ventilator use: binary indicator defined by ventilatory requirement

    - eICU: ventilator use defined using the 'Apache Patient Result' table and defining ventilator use as 'actual vent days' $\geq$ 0 within the prescribed offset of 1440
    - NTDB: ventilator use defined as 'VENTDAYS' with a duration >0
- Specific Comorbidities: Binary features were generated for the following comorbidities based on their shared availability in the source and target datasets:
    - hypertension
    - congestive heart failure (CHF)
    - cerebrovascular accident (CVA)
    - diabetes
    - chronic obstructive pulmonary disease (COPD)
    - myocardial infarction (MI)
    - renal disease
    - cancer
    - eICU: these comorbidities were obtained from the "past history" data file with the following dictionary
        - hypertension:
            - hypertension requiring treatment
        - CHF:
            - CHF - severity unknown
            - CHF - class I
            - CHF - class II
            - CHF - class III
            - CHF - class IV
        - CVA:
            - TIA(s) - within 6 months
            - TIA(s) - within 2 years
            - TIA(s) - within 5 years
            - TIA(s) – remote
            - TIA(s) - date unknown
            - Stroke - within 6 months
            - Stroke - within 2 years
            - Stroke - within 5 years
            - Stroke - remote
            - Stroke - date unknown
        - Diabetes:
            - Medication dependent diabetes
            - Insulin dependent diabetes
            - Non-medication dependent diabetes
        - COPD:
            - COPD - no limitations
            - COPD - moderate
            - COPD - severe
        - MI:
            - MI - within 6 months

        - MI - within 2 years
        - MI - within 5 years
        - MI – remote
        - MI - date unknown
    - Renal Disease:
        - Renal insufficiency - creatine 1-2
        - Renal insufficiency - creatinine 2-3
        - Renal insufficiency - creatinine 3-4
        - Renal insufficiency - creatinine 4-5
        - Renal insufficiency - baseline creatinine unknown
        - Renal failure - not currently dialyzed
        - Renal failure - hemodialysis
        - Renal failure - peritoneal dialysis
    - Cancer:
        - Bladder
        - Brain
        - Breast
        - Cancer
        - CML
        - Colon
        - Esophagus
        - Head and neck
        - Hodgkins's disease
        - Leukemia - other
        - Liver
        - Lung
        - Melanoma
        - Multiple
        - Multiple myeloma
        - Ovary
        - Pancreas - adenocarcinoma
        - Prostate
        - Testes
        - Uterus
- NTDB: these comorbidities were obtained from the "past history" data file with the following dictionary
    - Hypertension:
        - Hypertension requiring medication
    - Diabetes:
        - Diabetes mellitus
    - CHF:
        - Congestive heart failure
    - CVA:
        - CVA/residual neurological deficit

        - Cerebrovascular accident (CVA)
    - COPD:
        - Chronic Obstructive Pulmonary Disease (COPD)
    - MI:
        - History of myocardial infarction
    - Renal Disease:
        - Chronic renal failure
    - Cancer:
        - Disseminated cancer
        - Currently receiving chemotherapy for cancer
- Additionally, a comorbidity indicator variable was generated from the above dictionaries to indicate if a patient had a known history of any of the included comorbidities at the time of injury
- Nature of Injury:
    - eICU: These characteristics were extracted from the final element of the 'diagnosis' string and categorized as follows:
        - Fracture:
            - Bone fracture(s)
            - Fracture of skull
        - Internal organ:
            - Adrenal trauma
            - Bladder trauma
            - Blunt abdominal trauma
            - Bowel trauma
            - Cardiac injury – blunt
            - Diaphragmatic injury
            - Hepatic trauma
            - Intracranial injury
            - Lung trauma
            - Mesenteric trauma
            - Pancreatic trauma
            - Pneumomediastinum
            - Pneumothorax
            - Renal trauma
            - Retroperitoneal trauma
            - Spinal cord injury
            - Splenic trauma
            - Tension pneumothorax
            - Traumatic injury of esophagus
            - Ureteral trauma
        - Open Wounds:
            - Chest wall trauma
            - Penetrating abdominal trauma

        - Pellet injury from shotgun blast
        - Cardiac injury – penetrating
        - Airway trauma
- Sprain: sprain
- Dislocation: dislocation
- Blood vessels:
    - Traumatic injury to major vessel(s)
    - Extremity ischemia
    - Extremity compartment syndrome
    - Blunt vascular injury
    - Penetrating vascular trauma
    - Mesenteric ischemia
- Amputation: amputation
- Other: none

Additional Feature Coding for Baseline Models

The “All Features” baseline model allowed for the inclusion of additional features from the shared feature list, see Table 3.

- Temperature
- Respiratory Rate
- Supplemental Oxygen: a binary feature generated from ‘SUPPOXY’ variable. ‘Supplemental oxygen’ was coded as a 1 while ‘no supplemental oxygen,’ ‘not applicable,’ and ‘not recorded’ were coded as 0.
- Type of Transport:
    - Ground
    - Air
    - Other
- Trauma type: binary features for each of the following types:
    - Blunt
    - Penetrating
    - Burn
    - Other
    - Nan
- Trauma intent: binary features for each of the following intents:
    - Unintentional
    - Assault
    - Self-Inflicted
    - Other
    - NaN
- Trauma mechanism: binary features for each of the following mechanisms:
    - Fall
    - MVC:
        - MVT Occupant
        - MVT Motorcyclist
        - MVT Pedestrian
        - MVT Pedal cyclist
        - MVT Other
        - MVT Unspecified
    - Firearm
    - Struck by, against
    - Machinery
    - Transport:
        - Transport, other
        - Pedal cyclist, other
        - Pedestrian, other
    - Other:
        - Unspecified
        - Other specified and classifiable
        - Other specified, not otherwise classifiable

- Poisoning
- Natural/environmental, Bites and stings
- Natural/environmental, Other
- Suffocation
- Drowning/submersion
- Fire/flame
- Adverse effects, medical care
- Adverse effects, drugs
- Hot object/substance
- Overexertion

Missing Data

Missing data were approached at the facility level; therefore, the percentage of missingness by feature varied across facilities. While Figures 4 and 5 depict the percentage of missingness across the entire dataset, the percentage missing by facility can be found in Supplemental Table S1.

Please note that for the homogeneous TL approach, features were excluded only if the missingness exceeded 10% in the training data (source data combined with half of the target data). Knn imputation (*k*=5) was used when missingness was less than 10%.

- eICU:
    - Missing data are indicated by blank fields. We inserted "NaN" into these fields to prepare the data for imputation.
- NTDB:
    - indicators -1 and -2 were used in the NTDB for different purposes. A '-1' means 'not applicable' while a '-2' means 'not known or not recorded.' For features relevant to this study, such as age, sex, pulse, SBP, oxygen saturation, BMI, and indicator variables, data are expected to be available for all patients (e.g., every patient has a BMI). Therefore, -1 and -2 were treated as missing and substituted with "nan" with the following exceptions:
    - Ventilator use:
        - Missingness was treated as not requiring a ventilator (indicator =0)
    - Comorbidities:
        - Their absence was treated as not having those conditions (indicator =0)
    - Blood transfusion:
        - Absence was defined as not requiring transfusion (indicator = 0).

Test Dataset Preparation

Facility-specific NTDB data from 2016 were used as the test set for the TL-based model and all the baseline models. We prepared the data using the following steps:

- Subset patients to those seen in the 30 facilities we used to define the study cohort
- Removed the 45 patient encounters in which 'discharge disposition' was missing
- Subset cohort to those patients with 'ICUDAYS'>0
- Test cohort of 9045 patient encounters across 30 facilities

The feature development proceeded according to the steps outlined in the previous section, with a few exceptions. ICD 10 coding was implemented in October 2015. The NTDB 2016 data includes ICD 9 and 10 codes, however, D-codes and E-codes were less complete for ICD9 coding than for ICD10. Therefore, we used ICD 10 codes and translated them into ICD 9 code groupings to align with the study cohort (2014-2015). For example, 'cut/pierce' was added as an additional mechanism of injury in 2016. We grouped injuries using this mechanism as 'struck.'

Detailed Methods

Python version 3.8.5 was used for data management, analysis, and modeling. The following packages and libraries were used:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Joblib
- statsmodels.api

The following additional details accompany the development, training, and testing of the TL model, as well as the 'all features' and 'shared features' baselines:

- LASSO regularization was used to select the most important features in our feature sets
- LASSO used an L1 penalty and liblinear solver
- The selected features were then used in a logistic regression model, which was optimized on the training data using a liblinear solver and evaluating the following values for C: 0.1, 1.0, and 10.0
- The best performing model, according to F1 score, was then evaluated on the validation set with different thresholds for the positive class (0.05, 0.1, 0.15, 0.20, and 0.25)
- The optimal threshold was determined via bootstrapping (=100)
- This best performing model was then used on the test set for final model metrics, with bootstrapping (n=1000) to generate 95% confidence intervals for precision and recall

R, and the icdpicr package was used to map the ICD 9 codes found in the eICU data to the AIS scores.

ChatGPT was used to assist with coding and perform code reviews when errors occurred. Examples of prompts used include "I would like to generate a dot plot to represent many different entities according to a performance metric. I would like to generate multiple subplots, one for each entity. Please provide me with sample code" and "I would like to integrate LASSO regression for feature selection into my modeling code. Please provide me an example of how I would implement LASSO in this chunk of code."

Supplemental Tables & Figures

Table S1. NTDB percentage missingness, by facility, 2014-2015

	Shared Features							Additional NTDB Specific Features			
Facility	Age	Sex	Pulse	Syst BP	O2 Sats	GCS-c	GCS-t	BMI	ISS	Temp	Resp Rate
6547[tl]	3.68	0	0	0.46	1.61	0.92	0.69	11.03	0	22.53	2.30
2194[tl]	10.32	0	0	0	94.44	0	0	100.00	0	6.35	0
355[tl]	6.36	.19	0.58	0.39	52.56	20.15	0.96	58.53	0	3.66	1.16
6268[tl]	0	0	0.71	0.71	1.18	0.94	0.94	2.59	0	7.78	0.24
6552[tl]	4.99	0	0	0	0.50	0	0	4.99	0	5.49	0
6555[tl]	12.50	0	8.19	9.91	10.34	15.52	15.09	8.62	0	24.14	10.78
2005[tl]	3.30	0	1.27	3.18	3.30	5.72	3.81	100.00	0	8.26	3.18
2681[tl]	7.56	0	3.67	4.10	4.54	7.99	7.99	100.00	0	13.17	4.54
2705[tl]	6.60	0	0	0	0.47	1.42	0.47	7.55	0	2.36	1.42
2416[tl]	3.10	0	0.69	1.03	3.79	0.52	0.52	1.72	0	25.52	2.24
6242[tl]	1.87	0	0.47	0.47	1.40	1.87	1.87	0.47	0	16.36	2.80
2399[tl]	6.36	.46	0	0	1.52	0	0	8.18	0	31.36	0.15
2380[sh]	4.04	0	0	0.25	3.16	0	0	97.73	0	10.98	0
6362[all]	2.13	0	1.22	1.98	2.74	9.04	7.98	21.12	0	18.47	2.51
138[all]	4.83	0	0	0.09	0.55	0.55	0.46	7.75	0	2.10	0.55
2254[all]	5.38	0	0.64	0.51	2.69	0.64	0.64	100.00	0	9.36	0.51
2310[all]	0.71	0	0	0	0.95	0.48	0	85.04	0	6.89	0.24
5003[NTS]	1.68	0	11.76	12.61	18.49	14.29	14.29	55.46	0	22.69	14.29
2404[NTS]	3.90	0	0.46	1.38	2.29	1.38	1.15	29.59	0	7.80	1.61
2065[NTS]	4.91	0	0.64	1.07	2.56	2.99	2.56	100.00	0	17.95	1.71
2154[all]	2.60	0	0	0.37	0.74	0	0	17.84	0	2.97	0.37
6244[all]	4.88	0	7.01	7.64	9.98	16.56	16.56	2.97	0	25.90	7.64
358[all]	3.06	0	0	0	0.69	0	0	100.00	0	12.86	0.07
5004[all]	1.36	0	1.36	0.68	4.76	16.33	16.33	10.20	0	5.44	17.01
2001[all]	5.98	0	1.99	3.13	1.99	4.84	4.84	0.85	0	10.54	3.13
2199[all]	3.24	0	2.65	5.46	4.13	6.19	5.75	13.72	0	14.01	5.01
460[all]	0.92	0	0	0.35	0.35	1.50	0.69	2.88	0	1.38	0
315[NTS]	4.90	0	9.31	10.78	15.69	6.86	5.88	3.43	0	20.10	12.75
2680[all]	6.09	0	1.39	3.83	2.26	0.87	1.04	6.61	0	8.17	2.78
6131[all]	1.45	0	3.62	5.80	5.07	5.07	5.07	16.67	0	7.25	8.70

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best. *all* denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best.

Table S2. Characteristics of Treated Patients by Facility, 2014-2015

Facility	Patient Age (med)	Youngest Age Treated	Penetrating Trauma (%)	ISS (med)
6547[tl]	54	15	5.3	17
2194[tl]	62	18	1.6	17
355[tl]	51	<1	2.1	16
6268[tl]	56	15	5.1	17
6552[tl]	57	16	4.7	16
6555[tl]	69	14	4.7	17
2005[tl]	45	15	11.6	16
2681[tl]	61	16	2.8	16
2705[tl]	65	15	4.2	13
2416[tl]	56	14	7.6	14
6242[tl]	39	12	13.2	17
2399[tl]	62	16	7.6	17
2380[sh]	54	<1	5.9	17
6362[all]	56	<1	9.8	17
138[all]	52	14	5.5	17
2254[all]	55	15	5.6	17
2310[all]	42	10	11.6	13
5003[NTS]	37	<1	9.2	17
2404[NTS]	54	<1	9.2	16
2065[NTS]	51	<1	5.6	17
2154[all]	57	15	5.6	17
6244[all]	50	<1	8.0	14
358[all]	52	9	3.9	14
5004[all]	48	8	5.4	20
2001[all]	63	16	4.6	16
2199[all]	52	<1	10.2	17
460[all]	47	<1	5.8	16
315[NTS]	58	16	4.9	13
2680[all]	57	<1	4.3	10
6131[all]	40	13	6.5	22

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best. *all* denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best.

Table S3. Correlation Coefficients for a Subset of Predictors and the Outcome in Training Data

	Age	*Sex*	*Comorbidity-binary*	*Pulse*	*Systolic BP*	*Oxygen saturation*	*Vent-binary*	*GCS-motor*	*GCS-total*	*Blood transfusion*	*Internal Organ*	*Open wound*	*Fracture*
eICU	**0.091**	**-0.018**	**0.018**	**0.041**	**-0.070**	**-0.069**	**0.143**	**-0.387**	**-0.375**	**0.037**	**0.07**	**-0.02**	**-0.02**
6547^{tl}	0.127	-0.041	0.036	0.022	-0.082	0.018	0.209	-0.181	-0.213	0.155	-0.01	-0.08	-0.01
2194^{tl}	0.264	0.000	nan	-0.097	-0.008	0.180	0.046	-0.152	-0.151	0.253	-0.14	-0.02	0.04
355^{tl}	0.180	0.047	0.179	-0.024	0.034	0.046	nan	-0.389	-0.367	0.108	-0.08	-0.03	0.03
6268^{tl}	0.213	-0.034	0.077	0.012	0.111	-0.026	0.203	-0.248	-0.264	-0.055	0.05	-0.05	0.11
6552^{tl}	0.155	0.014	0.113	-0.029	-0.006	-0.261	0.281	-0.206	-0.227	0.066	0.04	-0.06	-0.03
6555^{tl}	0.059	-0.051	0.050	0.194	-0.122	-0.113	0.381	-0.317	-0.321	0.143	0.07	-0.06	0.06
2005^{tl}	0.224	0.009	0.138	-0.053	0.029	-0.169	0.261	-0.435	-0.439	0.202	0.09	-0.05	-0.01
2681^{tl}	0.195	0.060	0.097	-0.088	0.188	-0.019	0.132	-0.177	-0.193	nan	-0.07	0.04	0.04
2705^{tl}	0.035	-0.059	0.125	0.018	0.007	-0.019	0.275	-0.250	-0.269	0.117	-0.10	-0.06	-0.05
2416^{tl}	0.136	-0.053	0.163	-0.017	-0.036	-0.026	0.433	-0.417	-0.422	nan	0.06	-0.12	0.02
6242^{tl}	0.082	-0.033	0.055	0.059	0.056	-0.126	0.193	-0.198	-0.188	0.074	-0.01	0.02	0.00
2399^{tl}	0.186	-0.031	-0.024	0.048	0.028	-0.121	0.299	-0.332	-0.332	0.041	0.00	-0.05	0.04
2380^{sh}	0.157	-0.064	0.094	-0.049	-0.057	-0.161	0.286	-0.328	-0.302	nan	0.01	-0.02	-0.01
6362^{all}	0.141	-0.009	0.068	-0.058	-0.139	-0.258	0.275	-0.320	-0.313	0.046	0.04	-0.08	0.04
138^{all}	0.104	0.061	0.093	0.075	-0.100	-0.138	0.257	-0.377	-0.370	0.109	0.00	-0.04	-0.03
2254^{all}	0.143	0.046	0.056	-0.031	0.121	-0.047	0.069	-0.378	-0.381	nan	0.08	0.01	0.02
2310^{all}	0.087	0.006	-0.035	0.190	-0.029	-0.080	0.324	-0.450	-0.424	nan	0.02	-0.02	0.05
5003^{NTS}	0.438	0.289	0.327	-0.229	0.106	-0.171	0.192	-0.095	-0.073	-0.109	-0.04	0.00	-0.03
2404^{NTS}	0.159	-0.031	0.023	-0.015	-0.045	-0.143	0.264	-0.329	-0.355	0.116	-0.01	-0.05	0.05
2065^{NTS}	0.157	0.134	0.115	-0.007	0.007	-0.068	0.270	-0.267	-0.266	-0.004	0.06	-0.09	-0.05
2154^{all}	0.181	0.131	0.098	-0.029	-0.076	-0.041	0.258	-0.369	-0.387	nan	-0.07	-0.11	-0.03
6244^{all}	0.165	0.008	0.069	0.018	0.074	-0.045	0.233	-0.333	-0.348	-0.028	0.05	-0.03	0.04
358^{all}	0.134	-0.060	0.098	0.025	-0.038	-0.076	0.411	-0.459	-0.459	0.032	0.06	0.02	-0.04
5004^{all}	0.145	0.000	0.069	0.013	-0.037	-0.022	0.259	-0.245	-0.231	0.072	0.12	0.00	-0.15
2001^{all}	0.130	0.008	0.089	0.069	-0.028	-0.151	0.397	-0.440	-0.441	0.018	-0.09	-0.06	-0.01
2199^{all}	0.219	0.035	0.130	-0.017	-0.007	-0.090	0.415	-0.405	-0.402	0.127	0.03	-0.04	0.05
460^{all}	0.094	0.028	0.030	-0.040	-0.112	-0.041	0.275	-0.297	-0.282	0.101	-0.05	-0.04	-0.04
315^{NTS}	0.107	-0.029	0.042	0.022	0.097	-0.222	0.384	-0.391	-0.384	nan	0.17	0.03	-0.16
2680^{all}	0.136	0.065	0.091	-0.018	-0.001	-0.067	0.287	-0.260	-0.270	0.169	-0.02	0.00	-0.05
6131^{all}	0.150	-0.114	0.146	-0.033	-0.061	-0.302	0.191	-0.256	-0.241	0.267	-0.33	-0.10	-0.11

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best.

all denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best.

Table S4. Complete Performance Metrics, by Facility, across Modeling Approaches

	Train Data (2014/2015)		Test Data (2016)		Baseline Models									Aim1		
					BL-All			BL-Shared			BL-NTS			TL		
Facility	*n*	+ %	*n*	+%	*F1*	*PPV*	*Sensitivity*	*F1*	*PPV*	*Sensitivity*	*F1*	*PPV*	*Sensitivity*	*F1*	*PPV*	*Sensitivity*
6547[tl]	435	8.7	269	10.4	.207	.128	.536	.194	.117	.571	0.22	0.13	0.71	.379	.367	.393
2194[tl]	126	11.1	71	18.3	.31	.183	1.0	.31	.183	1.0	0.35	0.21	0.92	.571	.455	.769
355[tl]	1037	7.9	700	6.4	.194	.118	.533	.226	.138	.622	0.20	0.12	0.58	.353	.375	.333
6268[tl]	214	9.3	125	7.2	.143	.077	1.0	.136	.073	1.0	0.3	0.20	0.67	.414	.3	.667
6552[tl]	401	5.5	314	9.2	.32	.208	.69	.309	.196	.724	0.17	0.09	1.0	.436	.347	.586
6555[tl]	232	15.5	152	12.5	.222	.125	1.0	.222	.125	1.0	0.22	0.13	0.68	.295	.214	.474
2005[tl]	787	9.4	372	9.1	.359	.253	.618	.38	.288	.559	0.29	0.17	0.97	.5	.379	.735
2681[tl]	463	8.0	273	8.8	.191	.106	.958	.167	.092	.917	0.15	0.09	0.5	.246	.212	.292
2705[tl]	212	3.8	93	8.6	.168	.092	1.0	.165	.091	.875	0	0	0	.207	.143	.375
2416[tl]	580	9.8	274	6.6	.377	.286	.556	.248	.142	1.0	0.26	0.19	0.39	.286	.211	.444
6242[tl]	424	3.1	335	6.0	.209	.149	.35	.118	.067	.5	0.11	0.06	1.0	.216	.125	.8
2399[tl]	660	13.9	301	14.6	.333	.245	.523	.338	.256	.5	0.33	0.26	0.46	.38	.339	.432
2380[sh]	792	9.5	405	14.6	.376	.452	.322	.379	.5	.305	0.14	0.42	0.09	.362	.257	.61
6362[all]	471	8.9	274	13.1	.435	.316	.694	.332	.201	.944	0.35	0.25	0.58	.412	.328	.556
138[all]	1097	6.9	529	8.3	.392	.283	.636	.27	.179	.545	0.31	0.26	0.36	.370	.313	.455
2254[all]	780	7.8	393	8.4	.429	.405	.455	.221	.131	.697	0.36	0.27	0.55	.386	.458	.333
2310[all]	421	5.5	210	6.7	.357	.238	.714	.214	.214	.214	0.26	0.17	0.57	.316	.25	.429
5003[NTS]	119	12.6	52	9.6	.118	.069	.4	.105	.071	.2	0.18	0.1	1.0	.2	.12	.6
2404[NTS]	436	14.0	327	18.0	.443	.383	.525	.459	.367	.61	0.48	0.48	0.49	.389	.282	.627
2065[NTS]	468	13.5	250	14.8	.26	.149	1.0	.258	.148	1.0	0.37	0.27	0.62	.301	.306	.297
2154[all]	269	11.9	148	16.9	.41	.302	.64	.309	.182	1.0	0.27	0.31	0.24	.321	.232	.52
6244[all]	1316	10.4	894	9.2	.444	.358	.585	.327	.281	.390	0.37	0.27	0.59	.339	.337	.341
358[all]	1439	4.8	739	3.9	.468	.375	.621	.411	.359	.483	0.30	0.25	0.38	.333	.255	.483
5004[all]	147	8.8	55	16.4	.5	.571	.444	.281	.164	1.0	0.28	0.16	1.0	.353	.375	.333
2001[all]	351	12.5	177	13.0	.453	.4	.522	.281	.171	.783	0.4	0.34	0.48	.318	.215	.609
2199[all]	678	20.6	453	16.6	.502	.386	.72	.333	.202	.947	0.29	0.17	1.0	.346	.397	.307
460[all]	868	6.8	431	8.6	.382	.288	.568	.231	.133	.865	0.4	0.52	0.34	.274	.195	.459
315[NTS]	204	11.8	95	16.8	.491	.351	.813	.49	.364	.75	0.47	0.32	0.88	.313	.313	.313
2680[all]	575	10.4	276	12.7	.434	.375	.514	.287	.175	.8	0.29	0.2	0.54	.275	.178	.6
6131[all]	138	22.5	58	20.7	.316	.429	.25	.2	.25	.167	0.4	0.39	0.42	.143	.5	.083

tl denotes facilities in which the TL model performed best; *sh* denotes facilities in which the 'shared-features' baseline model performed best.

all denotes facilities in which the 'all-features' baseline model performed best; *NTS* denotes facilities in which the NTS baseline model performed best.

Primary Manuscript 2

Assessment of Heterogeneous Transfer Learning in Trauma Outcomes Prediction

Abstract

Importance:

In healthcare, obtaining or developing large datasets can be onerous and expensive. Most advanced analytics rely on big data to test hypotheses. Hospitals are often limited to small record sets or forced to use large datasets that are disconnected from local patterns of care. Transfer learning (TL) is a machine learning technique that augments datasets limited by small record sets with additional examples for ML models to learn while retaining the local context of the original data.

Objective:

To evaluate the efficacy of TL in predicting inpatient death following traumatic injury by augmenting hospital-specific datasets from the National Trauma Data Bank (NTDB) with a publicly available dataset of ICU patient stays (eICU CRD).

Study Design / Setting/ Participants:

Training data included trauma encounters between 2014 and 2015. From the eICU dataset, we selected encounters based on acute injury diagnoses (n=6,720). From the NTDB, 30 Level II trauma facilities were selected using stratified random sampling, and all ICU encounters within the selected facilities were included (n=16,140). We used three groups of features to build the model: features shared by the source and target data, those unique to the source, and those unique to the target. For features unique to one dataset, we used sampling techniques and k-nearest neighbors to perform imputation. Using logistic regression, we built and evaluated facility-specific models (n=30). We compared our results to previously developed models that used the same datasets and TL concepts, including:1) target-only, shared features baseline, target-only all features baseline, 3) target-only, restricted features baseline, and 4) a TL-based model restricted to the shared feature space.

Results:

The F1 scores for the heterogeneous TL model ranged from 0.1 to 0.51 with a mean of 0.32. The heterogeneous TL models scored comparably to or better than other baselines for nine of the 30 facilities (30%). We performed a post-hoc analysis of facility-specific models trained using all available level II trauma data in the NTDB. Without the incorporation of any source data, this model outperformed all other models in 16 of the 30 facilities (53.3%).

Conclusions:

Our results demonstrated inconsistent and marginal improvements in predictive performance when heterogeneous TL was used to develop facility-specific models. There are many potential reasons for these results with TL, which may include the following: dataset provenance, cohort selectivity, missing data prevalence, target data representation, and standardization of care processes being evaluated. We recommend further evaluation and development of criteria that can be used to determine the likelihood of TL conferring a predictive advantage.

Introduction

With the increasing availability of big data and advanced computational techniques, machine learning (ML) models are becoming common in medical settings. Generalizable models, which can be applied to any setting or patient cohort, are described as a goal of ML, yet sacrifices in performance are required to demonstrate such broad applicability. To date, there is a gap in modeling techniques that can learn overarching patterns from larger data sets that can better model facility-specific trends when incorporated with local data.

While these local effects may be important, they can be difficult to model with techniques that require large data sets. TL has been proposed as a way to retain the local salience of a problem under study while using the tooling of large data analytics. Transfer learning (TL), an ML-based approach, incorporates two disparate data sets: a source and a target. TL techniques learn patterns from the initial data set (the source) and apply relevant knowledge to the modeling task of a second data set (the target).[1] In this way, TL methods use the source data to learn a general phenomenon and the target data to glean the specificity of that trend locally. This approach addresses critiques of ML's reproducibility while acknowledging the need for local data to capture variability across health systems.

TL has shown promise in image recognition when using convolutional neural networks.[2] In binary classification tasks in the healthcare domain, TL has been used to predict inpatient mortality, death following surgery, and risk of *Clostridium difficile* infection during an inpatient stay.[3,4,5] The authors of these works share a similar motivation in preserving the underlying patterns and nuances of a local phenomenon that may be otherwise obscured by using a large database collected from dozens to hundreds of locations such as a registry.

Despite these examples demonstrating the advantage conferred on models using TL, the considerations for the use of this technique remain vague. In one review of TL, the authors state that the source and target data set should be 'of similar domains' to ensure TL has a positive effect on model performance.[6] The authors go on to suggest that the distribution of independent predictors (marginal probabilities), and distribution of independent predictors given the outcome (conditional probability) should be similar across datasets.[1,6,7] Finally, a shared precision in feature definition across datasets is recommended.[7] Beyond these guidelines, few resources exist to determine which data sets may be good candidates to employ in a TL evaluation.

To further evaluate the effect of TL, we applied it to trauma outcomes modeling. The study of traumatic injury in the United States has been greatly served by the continued development of the National Trauma Data Bank (NTDB). The American College of Surgeons maintains the NTDB, and it serves as a national registry for injured patients requiring hospital care across the United States.[8] Participating institutions, of which there are over 700, voluntarily submit data annually. Inpatient mortality following trauma has declined in incidence due to the standardization of trauma care guidelines and improved field and triage management.[9] This reduction in trauma mortality means that the difficulty of mortality prediction models increases for smaller volume trauma centers. By leveraging data from heterogeneous datasets and using TL techniques, the predictive capability within a smaller dataset can be augmented.

The main goal of this work is to determine whether heterogeneous TL positively affects trauma outcome prediction. To test our hypothesis, we will build and evaluate facility-specific TL models to determine the effect of this approach at the local level. We will augment the facility-specific NTDB data with eICU Collaborative Research Database data to form our training dataset. A heterogeneous TL approach means that the feature set will be composed of common features shared across both datasets and features unique to either the source data (eICU) or the target data (NTDB). We compared these results to baseline models formulated using traditional approaches to evaluate any added performance benefits incurred by TL.

Methods

1. Data Sources

Source Data:

The eICU Collaborative Research Database (eICU) is a de-identified electronic health record (EHRs) extract from patients admitted to intensive care units (ICU) at 208 facilities across the United States.[10] Encounters from the eICU database number over 200,000 and represent encounters from 2014-2015. This database is publicly available with approval for research purposes. This database includes granular data on patient demographics, admission diagnoses, vital sign measurements, laboratory results, and treatment plans.

Target Data:

The NTDB collects and aggregates data from US hospitals and then distributes datasets for research purposes.[8] The NTDB includes data such as demographics, injury severity, and physiologic data from trauma encounters. In the NTDB, we identified patients who received care

at ACS certified Level II trauma facilities between 2014 and 2016. The University of Washington Institutional Review Board determined that this study was exempt from IRB review because it did not meet the criteria for human subject research.

2. Study Cohort

Source Data:

We included all patient encounters in the eICU dataset for evaluation. The cohort was defined based on patient encounters associated with traumatic injuries. ICD-9 CM and ICD-10 CM coding were inconsistently available; therefore, string matching of the word "trauma" in the diagnosis was used to select the cohort (Figure 1). The excluded diagnoses for non-acute trauma-related admissions are detailed in *Supplement: eICU Cohort Determination.*

Target Data:

Thirty ACS Certified Level II trauma facilities and their associated patient care encounters were selected from the NTDB, with the only criterion being that the facility persisted in the NTDB across the years of study inclusion (2014 – 2016). The facilities were selected via stratified randomized sampling based on the bed size of the facility, with 60% of the facilities (n=18) having <400 patient beds and 40% of the facilities (n=12) having $\geq$ 400 patient beds, with a sampling ratio equivalent to the prevalence of large and small Level II facilities in the NTDB. Encounters where intensive care was required were included in the evaluation (Figure 2). Along with the source data, target encounters from 2014 and 2015 composed the training data while encounters from 2016 were retained for facility-specific model testing. Facility identifiers for each encounter were maintained so facility-specific data could be grouped and evaluated throughout the study. Data after 2016 was excluded from evaluation as persistent facility identifiers are no longer made publicly available in the NTDB for facility tracking over time.

3. Outcome & Predictors

Source Data:

The primary endpoint was death during the inpatient encounter. We excluded eICU encounters without a known discharge disposition. We evaluated all available data tables for feature extraction and consideration, including demographic information, comorbidities, vital signs, laboratory test data, medical therapies, physical exam findings, and home medications. While some tables provided data in numeric or categorical formats, others contained text strings. These strings were not free text but selected phrases in EHR drop-down menus. Binary

indicator variables were generated based on these text strings, such as 'acute distress,' an indicator variable derived from a text string indicating this finding on physical exam. We explicitly define these variables and the text strings from which we derived them in the *Supplement*. We used the first vital sign measurements and GCS scores in Table: Physical Exam to align with emergency room vitals and GCS scores in the NTDB. We defined the variable shock index (SI), a predictor of mortality in trauma patients, as the quotient of first available pulse and the systolic blood pressure.[11] Additional vital sign measurements were incorporated into the analysis based on their designation in Table: Physical Exam, including highest pulse, lowest systolic blood pressure, highest respiratory rate, lowest respiratory rate, and lowest oxygen saturation. We evaluated laboratory tests conducted within the first 24 hours of admission and intravenous infusions delivered over this time for additional features. Nursing documentation related to patient physical exam findings was considered for additional features. Admission medications were provided via free text fields with various names and spellings (e.g., metoprolol, Toprol XL, Lopressor). We used subject matter expertise to categorize these medications into 30 therapeutic classes (see *Supplement: Feature Development*). Of note, injury severity score (ISS), a commonly used descriptor of injury severity, was not available in the source data, nor could it be constructed from available data elements. Timestamps were unavailable in the source data; however, the data included time offsets based on patient admission time. We limited our features to data collected within the offset of 1440 (minutes within 24 hours).

Target Data:

Death and discharge to hospice care were used to define the primary endpoint.[12] We excluded NTDB encounters that did not have a discharge disposition. The NTDB data is available in a structured format across many tables. Features were extracted from the available variables, including demographic variables, physiologic features, mechanism and severity of injury variables (including ISS), and procedure details. (see *Supplement: Feature Development* for details on feature construction).

Additional variable groupings included:

- Vital signs and GCS scores obtained during emergency transport,
- Shock index based on emergency room vital signs,

- Patient payment class,
- Indicator of patient transfer,
- ICD 9 procedure codes,[13]
- Abbreviated injury scores per body region with a severity indicator if $\geq$ 3.

Across both source and target datasets, continuous predictors were evaluated for outliers or nonsensical values. The NTDB uses -1 and -2 to denote not applicable and unknown values, respectively.[14] Depending on the feature, we treated these values as 'missing' (details in the *Supplement: Missing Data*). Continuous predictors were scaled during the modeling and evaluation processes.

The source data was compared to facility specific target datasets to assess the degree of similarity between the datasets. Conditional probabilities were defined as the likelihood of a positive binary feature (e.g., blood transfusion = 1) among patients who died during the course of their care. These conditional probabilities were compared between source and target data sets. Fisher's exact test was used to evaluate the significance of the differences in the distributions.[15]

4. Information Gain Evaluation

We generated an array of binary variables using one-hot encoding for features with high cardinality. For example, within the target data, e-codes were grouped into 18 categories.[16] After one hot encoding, this resulted in 18 binary variables. This same approach was taken for procedure codes, source-derived home medications, and source-derived physical exam findings that were determined to be potentially predictive of severe disease and potential death by the study team. We expected this approach to result in significant sparsity among features. We evaluated the information gain associated with these features to reduce the feature space prior to modeling. Information gain is a feature selection technique that measures the entropy associated with a feature differentially across groups defined by the outcome.[17] Features with a higher entropy are considered less informative to the model. We used entropy-based feature selection to determine which variables should be retained for our modeling, with an entropy threshold of 0.9. For the target data, we evaluated information gain across the entire dataset, not at the level of the facility. Features with an entropy score of 0.9 and higher were not included in subsequent modeling.

5. Approach to Missing Data

The heterogeneous TL approach incorporates common features between the datasets with features unique to both the source and the target. Incorporating these dataset specific features required mapping the datasets to a shared feature space.[18] This involved creating new variables to correspond to the unique variables in each set. For example, *ISS* is a feature unique to the target dataset, the NTDB. We generated a variable *ISS* for the eICU data that we imputed with NaN values. Similarly, the source data had a feature for *highest pulse* that was not available in the target data. We generated a *highest pulse* variable for the target data and imputed the variable with NaN values. Once we generated new variables for each of the unique variables, we were able to merge our source and target data sets. Our datasets now had three types of features: 1) common features, 2) source specific features, and 3) target specific features.

Common Features

There were 29 features in common between the source and target datasets. When facility specific NTDB datasets were combined with the eICU data, feature level missingness among these common features was assessed. These missing values were imputed using k-nearest neighbor imputation (k=3). The imputer used instances from the source and target datasets to inform the imputation.

Source & Target Features

Features uniquely derived from the source (n=13) and target (n=55) will have large degrees of missingness due to their method of development. The prevalence of missing data will vary depending on the size of the target training data (facility-specific). For example, the largest hospital in our target dataset has over 1200 encounters in the training data. When combined with the source data, the proportion of the target data is ~ 18%. For these facilities, source-specific features will have missing values ~18% of the time (the prevalence of the target data, when source specific features are 'NaN'). For these same facilities, target specific features will have missing values in ~82% of instances (1- the prevalence of the target data, when target specific features are 'NaN'). Among the facilities with smaller record set sizes, this becomes increasingly imbalanced.

When imputing source-specific features into rows from the target record set, we used k-nearest neighbors imputation (k=3). Records from the source data only were used to inform the imputation of target rows. Given the imbalance in record set size, we sampled the source record

set via the distribution of a key common feature, 'ventilator use'. This reduced source record set is then used to identify the nearest neighbors for feature imputation. This is similar to a method proposed by Wu et al, termed *feature creation imputation*.[18]

When imputing target-specific feature into rows from the source record set, we must take a different approach. The imbalance in record set sizes is so vast, that a feature creation imputation approach would necessitate replicating the target record set size 5 to 55 times. To avoid the overfitting that would result from that replication of data, we use the mean of that value based on the target data to impute into the source rows. The effect of this imputation approach will be evaluated by comparing the Pearson's correlation coefficient among binary variables before and after imputation occurs.[19]

6. Facilities

To characterize the contexts in which TL is an effective modeling approach, we described the included NTDB facilities according to the size of their trauma patient population. Proxies for resource availability included the number of ICU trauma beds and the number of trauma surgeons on staff. The patient cohort receiving trauma care at each facility was characterized by age, penetrating trauma prevalence, and injury severity. The difference in the prevalence of serious injury was evaluated using the Chi^2 test and significance was defined if the primary admission cohort differed from the interfacility transfer cohort at the level of $p \leq .05$.

7. Statistical Analysis

Model Development

We developed a heterogeneous TL model using logistic regression and a holdout set optimized for the F1 score. Coefficients from multivariate analyses were estimated using logistic regression and were used to determine the strength of the predictor.[20] Given our interest in evaluating the model's performance on facility-specific target data, the holdout set contained only NTDB encounter data from the facility whose data was used in training. While a variety of ML algorithms could have been used in this evaluation, guidance in the literature recommends the use of less complex ML algorithms with TL techniques. For example, Curth *et al.* used logistic regression and gradient boosted classifiers to highlight the relative effect of TL on model performance.[7] Whereas other algorithms have numerous hyperparameters that can be tuned to improve model performance, logistic regression does not. Therefore, the relative effect of TL on

model performance may be more evident. In keeping with this guidance, we will only evaluate logistic regression-based models in this analysis.

As shown in Figure 3, the modeling was performed at the facility level. Patient records from each NTDB facility were combined with the eICU dataset to form the study cohort. Half of each facility's data were selected as the holdout set. The training set was then used to set the model parameters. Given the relatively small data size and to avoid overfitting, we constrained our feature sets by using LASSO.[20] LASSO, or Least Absolute Shrinkage and Selection Operator, is a regression-based approach to feature selection that aims to reduce redundancy across included features. LASSO does this by reducing the coefficients, often to zero, of features that do not contribute additional information to a predictive model. Following the entropy-based feature selection step, we executed additional feature selection during the facility-specific modeling process. While the potential features for inclusion were consistent across facilities, the final features included in the model differed according to their LASSO coefficients. After each model was evaluated and tuned on the validation set, the fit model was evaluated using the 2016 data from each facility, which served as the test set.

We bootstrapped our models and evaluated the average F1 score across 1000 model iterations. Bootstrapping involved sampling with replacement of the test set to better estimate the variability and performance of our models.[22] We used Python 3.8 for data management, model building, analysis, and evaluation.[23] Details of the Python packages used in these processes are available in the *Supplement: Detailed Methods.* ChatGPT was used throughout the exploratory data analysis and model development processes to assist with Python coding and code error mitigation.[24] No data were passed on to ChatGPT, nor did ChatGPT provide any analysis of the results.

8. Model Evaluation: Discrimination

For each of the 30 selected facilities, we built a model using features common across both data sets as well as features unique to the source and target data sets (Figure 3). In each modeling stage, mortality during the inpatient encounter was defined as the positive class and labeled as $f(x) = 1$. Patient survival to discharge was defined as a negative class, labeled as $f(x) = 0$. The models output a probability for each patient encounter that was scored. Based on thresholds defined upon evaluation, the probability was translated into either 1 (for inpatient death) or 0 (for survival to discharge). After the encounters were scored, the predicted label was compared to the ground truth for each encounter. We evaluated all models using the F1 score which is the harmonic mean of sensitivity and the positive predictive value.[25]

9. Model to Baseline Comparison

In our prior work we evaluated the performance of homogenous TL in predicting trauma outcomes. To determine if homogeneous TL improved predictive performance, we developed baseline models for comparison. We built three baseline models for each facility included in the evaluation. Each model was defined as follows:

- "shared features" baseline model: This model's training data contained facility-specific target only data. The feature set was limited to those features the target and source data held in common.
- "all features" baseline model: This model's training data contained facility-specific target only data. Any available features in the NTDB were allowed for inclusion .
- "revised trauma score" baseline: This model's training data contained facility-specific target only data. We used a feature set limited to age, injury severity score, systolic blood pressure, GCS total and respiratory rate to represent the Revised Trauma Score.[26]

Like the heterogeneous TL approach, the source and target data sets are combined to comprise the training data in homogeneous TL. The difference in the two approaches lies in the feature space. While the heterogeneous approach uses the union of the source and target data, the homogeneous approach is restricted to the features shared between the data sets.
We compared our heterogeneous TL models to each of the 4 models previously developed for each facility (the three baselines and the homogeneous TL model). We compared these heterogeneous TL models to facility-only baseline models and a homogeneous TL model.
The relative performance of a TL model compared to another modeling approach can be considered in terms of the 'Transfer Gap', that is the absolute or relative difference in performance between the two methods.[26] We compared the impact of heterogeneous TL at the facility level by comparing the best-scoring baseline model (via the F1 score) with the heterogeneous TL model.

Results

1. Participant Characteristics

Source Data:

After removing 85 (1.2%) patient encounters for which no outcome could be ascertained, a total of 6,720 patient encounters composed the study's source data from the eICU, of which 499

patients died (7%). Among the source data encounters, the median age of surviving patients was 58 years, whereas the median age of patients who died was 71. Among surviving patients, 63% were male; 2/3 of patients who died were male. This cohort is further described in Table 1.

Target Data:

Across the 30 included facilities, 16,140 patients received ICU care and met our inclusion criteria to compose the study cohort, of which 1518 died (8%). Among those classified as deceased in our evaluation, 199 were discharged for hospice care (13%). The median age of surviving patients was 51 years, whereas the median age of patients who died was 67. Among surviving patients, 67% were male; 65% of patients who died were male. This cohort is further described in Table 1 with additional trauma-related characteristics described in Table 2.

Test Data:

NTDB data from 2016 served as our test data. This included 9,045 patient encounters, of which 921 died (10.2%). Among surviving patients, the median age was 44 years, while the median age of patients that died was 46. Over 60% patients in the test cohort were male. This cohort is further described in Table 1 with additional trauma-related characteristics described in Table 2.

2. Facility-Level Characteristics

The 30 trauma facilities that were selected via stratified random sampling are listed in *Supplement: Table 1*. Twelve facilities were "large" with $\geq$ 400 inpatient beds, and 18 "small" with <400 inpatient beds. Large hospitals averaged 442 trauma admissions requiring ICU care per year, whereas small hospitals averaged 154. The mortality rate among these patients varied between facilities, with a minimum of 3.1% and a maximum of 22.5%. The mean mortality rate among large facilities was 9.9% and the mean mortality rate among smaller facilities was 10.4%. Within each facility, the distributions of characteristics in the NTDB training and testing cohort were compared within facilities using the Wald test. There was no significant difference ($p \leq .05$) in the distribution of data between each of the 30 facilities across the characteristics available. These facilities are described in *Supplement: Table 2*.

3. Variable Selection & Variable Importance

We used information gain as a variable selection technique before training and fitting the models. Many of these resulted from binary features that were derived from categorical features with high cardinality. Based on the 0.9 threshold for entropy, we retained the following features from the source data tables: elevated PTT indicator, elevated INR indicator, receipt of FFP, acute distress, treatment with inotropes, treatment with vasopressors, treatment with blood products, and vasoactive infusions. None of the features generated from home medication classes generated an entropy score low enough be included. From the target data we evaluated 18 categories of external cause of injury codes (e-codes) and 18 categories of procedure codes with this same method. Based on the results, we retained three external cause of injury groupings (‘other road vehicle’, ‘self-inflicted’, ‘assault’) and three procedure code groupings ('nervous system procedures', 'ocular procedures', ‘ear, nose, and throat procedures’). These retained variables were added to the feature space for modeling are summarized in Table 3. The entropy values are available in *Supplement: Tables 3, 4, and 5.*

Conditional probability distributions were compared between the source data and the target datasets, pre-imputation among features in the shared-feature category. The degree of alignment between source and target data sets is described in *Supplement: Table 6.*

4. Data Missingness and Imputation

We compared training data distributions pre- and post- imputation for each facility-specific dataset. The correlations of binary variables to the outcome from the shared feature set and the target specific feature set were evaluated. The pre- and post-imputation correlation coefficients were compared. An example of this evaluation is shown in Figure 4, with additional examples available in the *Supplement: Figure 1*. *C*omplete calculations are available upon request.

5. Model Performance

Overall, the heterogeneous TL models performed poorly, with F1 scores ranging from 0.1 to 0.51 with the facility-specific results shown in Table 4. The mean F1 score across all the facilities was 0.32. The TL approach performed better among facilities with smaller record sets (mean F1 of 35.2) than among those with larger record sets (mean F1 of 28.2). The highest-performing facilities generated F1 scores of 0.50 (0.49-0.51, n=5). The patient cohorts at these five facilities were slightly younger (mean age 49 vs 54.3) and had higher penetrating trauma rates (8.2% vs 6%) than the remaining facilities. We evaluated the correlation between the proportion of training data derived from the target and the F1 score. There proved to be no

correlation with a Pearson's correlation coefficient of -.28 indicating a weak and negative correlation between increasing target data proportion and F1 score. This relationship is shown in Figure 4.

6. Comparison to Baselines and Homogeneous TL Approach

Building on our prior unpublished evaluation of TL on a common data set (homogeneous TL), we can compare the heterogeneous TL models to the earlier developed baselines and the homogeneous TL models. The F1 scores for all facilities and models are available in Table 5. The results can be summarized as follows:

- All models had a similar distribution of F1 scores (Figure 5) with the following mean F1 scores across all facilities:
    - "all" features baseline: 0.34
    - "common features" baseline: 0.27
    - "RTS features" baseline: 0.29
    - Homogenous TL: 0.33
    - Heterogeneous TL: 0.32
- The "all-features" model provided the highest performance (or tied) across 11 facilities (36%). Among these facilities, the all-features model averaged an F1 score of 0.45.
- The homogeneous TL model scored highest (or tied) among 7 facilities with an average score of 0.37.
- The heterogeneous TL model scored highest (or tied) among 9 of the facilities with an average F1 score of 0.45.
- Table 5 shows the comparison of heterogeneous TL scores compared to best performing baselines. The relative and absolute transfer gap is provided. An absolute negative transfer gap of > 0.02 by F1 score occurred in half of facilities (n=15). A relative negative transfer gap of >2% by F1 score occurred in 17 of 30 facilities (56.7%).

7. Post Hoc Analysis

After initial model results and comparisons to baseline proved underwhelming, we retrained our models and evaluated the impact of two alterations. First, we optimized our models for AUC instead of F1, to see if model performance varied when using a threshold-agnostic measure. To compare these results to baseline, we also evaluated those models when optimized for AUC. Second, we increased the relative influence of the target data within the training set by

weighting it five times the base weight. Increasing the weight of the target data is a common approach in TL and Desautels et al. determined a weight of 5x to be most beneficial to their performance metrics.[3] Overall, AUC scores varied from 0.5 to 0.89 with a mean of 0.74. Compared to baselines, the heterogeneous TL model performed relatively better when using AUC to rank performance. The TL model outscored all models in 11 facilities (37%). When evaluating models with weighted target data in the training set, F1 scores improved, on average, by 17%, however, this average increase is largely due to a few outlier facilities. The median change in F1 score with this approach is 0. When ranking modeling approaches compared to baselines, the TL model with weighted target data outperformed all baselines in 4 facilities (13%). These full metrics for these additional analyses are available upon request.

As part of our ad hoc analysis, we also compared heterogeneous TL models to a registry-like model built on additional data from the NTDB. This approach differed from the 'all features' baseline model in the size of the training data. Instead of using only facility specific training data as we did for other baselines, we used all ICU patient encounters at ACS verified level II trauma centers in 2014 and 2015 in the NTDB. In effect, we are trialing the effect of using a large registry without inducing a limited record set. Compared to other approaches, the exploratory 'all NTDB' approach performed best overall, with a mean F1 score of 0.40 with the facility-specific results shown in Table 5. The 'all NTDB' was the most consistent approach, generating the top performing F1 score in 16 of the 30 facilities. When comparing the heterogeneous TL approaches to this 'all NTDB' model, the relative negative transfer gap (the difference in F1 scores divided by the best-performing baseline) ranged from 1% to 73% across 24 facilities, with an average of 30%, demonstrating that baseline approaches significantly outperformed the heterogeneous TL approach. There were 6 facilities wherein the heterogeneous TL model outperformed the 'all NTDB' model. When comparing the characteristics of these facilities, the record sets were more limited (average of 439 vs. 563) as was the size of the positive class (mean of 46 vs 52).

Discussion

In our analysis, predicting mortality following trauma using heterogeneous TL methods failed to consistently improve upon the classification performance of other approaches. Heterogeneous TL models tended to perform better than baseline methods and other approaches among facilities with smaller record sets and reduced positive class sizes. While the heterogeneous approach expanded the included feature sets to include source and target specific features,

their inclusion did not significantly improve model performance compared to the homogeneous TL approach or to baselines. The approach to the missing data incurred by including source and target specific features represents a novel combination of methods. However, this imputation approach generally weakened the correlation between target-specific features and the outcome. Finally, the 'all NTDB' comparison model proved to be the most consistent approach to predicting mortality following trauma at the facility-level, demonstrating the value of dataset provenance.

Conclusion

Our modeling efforts with heterogeneous TL resulted in facility-specific models with inconsistent and underwhelming performance, measured via F1 score. These models tended to perform better when the proportion of target data in the training set was reduced, yet this correlation was weak. This approach provides a useful perspective when considering the use of TL and the impact of a combined feature set. Further work evaluating additional imputation methods that retain the information inherent in target specific features may be valuable.

References

1. Pan SJ, Yang Q. A survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2010;22(10):1345-1359. doi:10.1109/tkde.2009.191

2. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC medical imaging. 2022 Apr 13;22(1):69.

3. Desautels T, Calvert J, Hoffman J, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical Informatics Insights*. 2017;9:117822261771299. doi:10.1177/1178222617712994

4. Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. *2012 IEEE 12th International Conference on Data Mining Workshops*. Published online 2012. doi:10.1109/icdmw.2012.93

5. Wiens J, Guttag J, Horvitz E. A study in Transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*. 2014;21(4):699-706. doi:10.1136/amiajnl-2013-002162

6. Weiss K, Khoshgoftaar TM, Wang D. A survey of Transfer Learning. *Journal of Big Data*. 2016;3(1). doi:10.1186/s40537-016-0043-6

7. Curth A, Thoral P, van den Wildenberg W, et al. Transferring clinical prediction models across hospitals and Electronic Health Record Systems. *Machine Learning and Knowledge Discovery in Databases*. Published online 2020:605-621. doi:10.1007/978-3-030-43823-4_48

8. Hashmi ZG, Kaji AH, Nathens AB. Practical guide to surgical data sets: National Trauma Data Bank (NTDB). *JAMA Surgery*. 2018;153(9):852. doi:10.1001/jamasurg.2018.0483

9. Sakran JV, Jehan F, Joseph B. Trauma systems: standardization and regionalization of care improve quality of care. Current Trauma Reports. 2018 Mar;4:39-47.

10. 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). doi:10.1038/sdata.2018.178

11. Cannon CM, Braxton CC, Kling-Smith M, Mahnken JD, Carlton E, Moncure M. Utility of the shock index in predicting mortality in traumatically injured patients. Journal of Trauma and Acute Care Surgery. 2009 Dec 1;67(6):1426-30.

12. Kozar RA, Holcomb JB, Xiong W, Nathens AB. Are all deaths recorded equally? the impact of hospice care on risk-adjusted mortality. *Journal of Trauma and Acute Care Surgery*. 2014;76(3):634-641. doi:10.1097/ta.0000000000000130

13. ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles. Centers for Medicare & Medicaid Services. 2023. Available at: cms.govhttps://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-9-cm-diagnosis-procedure-codes-abbreviated-and-full-code-titles

14. American College of Surgeons. National Trauma Data Bank National Trauma Data Standard. Data Dictionary. 2016 Admissions. Available at: https://www.facs.org/~/media/files/quality%20programs/trauma/ntdb/ntds/data%20dictionaries/ntds%20data%20dictionary%202016.ashx.

15. Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restorative dentistry & endodontics. 2017 May 1;42(2):152-5.

16. ICD Injury Codes and Matrices. Centers for Disease Control and Prevention. 2021. Available at: https://www.cdc.gov/nchs/injury/injury_matrices.htm.

17. Information Gain and Mutual Information for Machine Learning. Machine Learning Mastery. 2020. Available at: https://machinelearningmastery.com/information-gain-and-mutual-information/#:~:text=Information%20gain%20is%20the%20reduction,before%20and%20after%20a%20transformation.

18. Wu X, Khorshidi HA, Aickelin U, Edib Z, Peate M. Transfer learning to enhance amenorrhea status prediction in cancer and fertility data with missing values. Artificial Intelligence: Applications in Healthcare Delivery. 2020 Dec 2:233.

19. Obilor EI, Amadi EC. Test for significance of Pearson's correlation coefficient. International Journal of Innovative Mathematics, Statistics & Energy Policies. 2018 Jan;6(1):11-23.

20. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: An overview. *Journal of Thoracic Disease*. 2019;11(S4). doi:10.21037/jtd.2019.01.25

21. Muthukrishnan R, Rohini R. Lasso: A feature selection technique in predictive modeling for Machine Learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. Published online 2016. doi:10.1109/icaca.2016.7887916

22. Nakatsu RT. An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems*. 2021;36(3):51-57. doi:10.1109/mis.2020.2978066

23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*. 2011;12:2825-30.

24. OpenAI. ChatGPT, 2023. [Large Language Model]

25. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models. *JAMA*. 2017;318(14):1377. doi:10.1001/jama.2017.12126

26. Jeong JH, Park YJ, Kim DH, et al. The New Trauma Score (NTS): A modification of the revised trauma score for better trauma mortality prediction. *BMC Surgery*. 2017;17(1). doi:10.1186/s12893-017-0272-4

27. Wang Z, Dai Z, Poczos B, Carbonell J. Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2019. doi:10.1109/cvpr.2019.01155

28. Committee on Trauma American College of Surgeons. Resources for Optimal Care of the Injured Patient: American College of Surgeons, 2014. American College of Surgeons website.

Available at: https://www.facs.org/-/media/files/quality-programs/trauma/vrc-resources/resources-for-optimal-care.ashx

29. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951 Mar 1;22(1):79-86

30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-57.

31. Brown JB, Watson GA, Forsythe RM, er al. American College of Surgeons trauma center verification versus state designation: are Level II centers slipping through the cracks?. The journal of trauma and acute care surgery. 2013 Jul;75(1):44.

Tables

Table 1. eICU and NTDB Training Cohort Descriptions (2014, 2015).

	eICU				NTDB							
	Study Cohort (2014, 2015)								Test Cohort (2016)			
	Survived (n=6221)		Died (n=499)		Survived (n=14622)		Died (n=1518)		Survived (n=8124)		Died (n=921)	
Feature	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Age	58	22.5	71	23.8	51	22.9	67	22.7	44	37.1	46	43.5
Sex	.63	.483	.66	.473	.67	.469	65%	.476	.66	.475	.63	.483
Syst BP	128	25	123	31	137	28.5	138	41.9	136	33.6	138	47.9
Pulse	85	19	87	24	88	22.4	88	28.8	88	24.8	86	31.65
Oxy Sat	97	3.7	96	7.4	97	4.6	95	9.1	93.8	17.4	88.6	25.5
Vent Bin	.595	.491	.861	.346	.276	.447	.724	.447	.27	.444	.746	.435
GCS eye	3.59	.9	2.26	1.4	3.6	.99	2.4	1.4	3.44	1.31	2.3	1.67
GCS verb	4.2	1.4	2.38	1.7	4.22	1.34	2.6	1.8	4.08	1.66	2.52	2.01
GCS motor	5.6	1.2	3.52	2.2	5.4	1.45	3.5	2.3	5.24	1.84	3.39	2.53
GSC total	13.39	3.18	8.17	5.04	13.2	3.6	8.6	5.4	12.93	4.14	8.38	5.7
Blood trans	.024	.152	4.6	.21	.047	.211	.105	.306	.007	.085	.014	.118
Comorbid	.537	.5	.571	.5	.392	.488	.549	.498	.386	.487	.537	.499
Hypertension	.33	.47	.37	.48	.31	.46	.43	.495	.312	.464	.436	.496
Diabetes	.144	.35	.15	.35	.12	.32	.16	.364	.132	.338	.15	.357
CHF	.058	.23	.094	.292	.03	.17	.07	.259	.032	.176	.089	.285
CVA	.057	.23	.068	.252	.02	.15	.05	.225	.026	.158	.064	.245
COPD	0	0	0	0	0	0	0	0	.058	.234	.1	.3
MI	.043	.20	.056	.23	.01	.11	.018	.135	.009	.094	.014	.118
Renal	.041	.197	.068	.252	.01	.1	.028	.166	.016	.126	.051	.22
Cancer	.06	.24	.08	.27	.007	.08	.020	.141	.009	.094	.02	.139
Internal organ	.577	.49	.709	.45	.70	.458	.713	.452	.736	.441	.759	.428
Open wounds	.08	.28	.07	.26	.367	.482	.307	.461	.356	.479	.295	.456
Fracture	.47	.5	.44	.5	.610	.488	.605	.49	.628	.483	.631	.482
Blood vessels	.02	.14	.016	.126	.044	.205	.06	.237	.05	.217	.05	.218
Sprains	0	.03	0	0	.059	.236	.02	.139	.067	.249	.028	.166
Dislocation	.0106	.10	.004	.063	.048	.214	.035	.184	.063	.243	.056	.231
Amputation	0	0	0	0	.014	.119	.007	.085	.004	.062	.011	.104
Other	.065	.25	.144	.35	0	0	0	0	.027	.161	.017	.131

Table 2. Description of Additional NTDB Characteristics

	Study Cohort (2014, 2015)				Test Cohort (2016)			
	Survived (n=14622)		Died (n=1518)		Survived(n=8124)		Died (n=921)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ISS	10	10	21	14.4	12	9.27	29	13.86
Type of Injury								
Blunt	.863	.344	.855	.352	.879	.327	.866	.34
Penetrating	.063	.243	.072	.258	.066	.248	.064	.245
Burn	.001	.038	.003	.057	.001	.037	.001	.033
Other	.031	.174	.036	.185	.014	.117	.011	.104
Mechanism of Injury								
Fall	.403	.49	.552	.497	.412	.492	.561	.496
MVC	.326	.469	.25	.433	.331	.471	.239	.427
Firearm	.033	.178	.063	.242	.036	.186	.055	.229
Struck	.059	.235	.022	.146	.087	.282	.023	.149
Machinery	.003	.057	.003	.051	.002	.041	0	0
Transport	.072	.259	.028	.051	.058	.234	.033	.18
Other	.028	.165	.038	.192	.04	.196	.058	.233
Intent of Injury								
Unintentional	.843	.364	.856	.351	.847	.360	.86	.347
Assault	.089	.285	.047	.211	.089	.285	.033	.18
Self-Inflicted	.019	.137	.053	.225	.017	.13	.04	.196
Undetermined	.005	.07	.006	.077	.007	.083	.01	.098
Other	.003	.054	.003	.057	.027	.161	.017	.131

Table 3. Features Used in Heterogeneous TL Approach

Variable Type	Source (n=13)*	Target (n=55)	Common (n=29)
Admission Details[1]	0	13	0
Patient Demographics[2]	0	4	2
Patient History[3]	0	0	9
Home Medications[4]	0	0	0
Injury Details[5]	0	30	8
Physical Exam[6]	1	0	0
Vital Signs[7]	5	4	8
Laboratory Results[8]	0	0	0
Initial Treatment[9]	7	4	2

1 includes features such as EMS vitals and GCS scores, mode of transport to the ED,
2 includes features such as age, sex, payor type,
3 includes binary indicators of cancer, heart failure, myocardial infarction, renal disease, hypertension, diabetes, stroke, and COPD,
4 while these features were developed, none of them were retained for modeling following an evaluation of information gain,
5 includes features indicating type of injury, mechanism of injury, and severity of injury,
6 includes binary indicator of acute distress,
7 includes ED based vital signs and highest pulse, lowest blood pressure, highest respiratory rate, lowest respiratory rate, and lowest oxygen saturation,
8 includes binary indicators for elevated INR and elevated PTT,
9 includes binary indicators for treatment types provided including blood products, vasopressors, and procedure groupings.

Table 4. Facility F1 Scores Across All Modeling Types

Facility	All Features BL	Common Features BL	RTS BL	Homogeneous TL	Heterogeneous TL	ALL NTDB
6552	0.32	0.31	0.17	0.44	0.51	.34
5003	0.12	0.11	0.18	0.14	0.5	.33
6547	0.21	0.19	0.22	0.38	0.5	.38
2005	0.36	0.38	0.29	0.5	0.49	.5
2199	0.50	0.33	0.35	0.35	0.49	.44
2154	0.41	0.31	0.27	0.32	0.42	.48
6131	0.32	0.2	0.4	0.14	0.4	.47
315	0.49	0.49	0.47	0.31	0.39	.53
2404	0.44	0.46	0.48	0.39	0.39	.58
358	0.47	0.41	0.30	0.33	0.36	.42
2194	0.31	0.31	0.35	0.57	0.36	.54
2705	0.17	0.17	0	0.21	0.36	.46
6362	0.44	0.33	0.35	0.41	0.36	.51
460	0.38	0.23	0.4	0.27	0.34	.41
2001	0.45	0.28	0.4	0.32	0.33	.44
6555	0.22	0.22	0.22	0.3	0.33	.33
138	0.39	0.27	0.31	0.37	0.29	.33
6268	0.14	0.14	0.3	0.41	0.29	.31
2399	0.33	0.34	0.33	0.34	0.28	.42
2065	0.26	0.26	0.37	0.30	0.26	.41
2416	0.19	0.24	0.26	0.29	0.26	.39
2680	0.43	0.29	0.29	0.28	0.26	.37
2254	0.43	0.22	0.36	0.39	0.22	.32
2681	0.19	0.17	0.15	0.25	0.21	.25
6242	0.21	0.12	0.11	0.22	0.21	.38
6244	0.44	0.33	0.37	0.34	0.21	.38
5004	0.5	0.28	0.28	0.35	0.18	.35
2380	0.38	0.38	0.14	0.36	0.12	.44
2310	0.36	0.21	0.26	0.32	0.11	.38
355	0.19	0.23	0.20	0.35	.077	.04

Table 5. Relative and Absolute Transfer Gaps from Heterogeneous TL Model to Best Performing Baselines

Facility	TL - Combined Vars	Best Baseline	Absolute Gap from Best Baseline	Relative Gap from Best Baseline	ALL NTDB	Absolute Gap from All NTDB	Relative Gap from All NTDB
6552	0.51	0.32	0.19	59%	0.34	0.17	51%
6547	0.5	0.21	0.29	142%	0.38	0.12	31%
5003	0.5	0.18	0.33	186%	0.33	0.17	50%
2005	0.49	0.38	0.11	29%	0.5	-0.01	-1%
2199	0.49	0.5	-0.01	-2%	0.44	0.05	10%
2154	0.42	0.41	0.01	2%	0.48	-0.06	-12%
6131	0.4	0.4	0	0%	0.47	-0.07	-15%
2404	0.39	0.48	-0.09	-18%	0.58	-0.19	-33%
315	0.39	0.5	-0.1	-20%	0.53	-0.14	-27%
2705	0.36	0.17	0.20	117%	0.46	-0.10	-21%
2194	0.36	0.35	0.01	3%	0.54	-0.18	-33%
6362	0.36	0.44	-0.08	-17%	0.51	-0.15	-29%
358	0.36	0.47	-0.11	-23%	0.42	-0.06	-14%
460	0.34	0.4	-0.06	-15%	0.41	-0.07	-16%
2001	0.33	0.45	-0.12	-26%	0.44	-0.11	-24%
6555	0.33	0.22	0.11	51%	0.33	0.00	0%
138	0.29	0.39	-0.10	-26%	0.33	-0.04	-13%
6268	0.29	0.3	-0.01	-5%	0.31	-0.03	-9%
2399	0.28	0.34	-0.06	-17%	0.42	-0.14	-34%
2680	0.26	0.43	-0.17	-40%	0.37	-0.11	-29%
2065	0.26	0.37	-0.11	-30%	0.41	-0.15	-37%
2416	0.26	0.26	0.00	1%	0.39	-0.13	-34%
2254	0.22	0.43	-0.21	-49%	0.32	-0.10	-31%
6242	0.21	0.21	0.01	2%	0.38	-0.16	-43%
6244	0.21	0.44	-0.23	-53%	0.38	-0.17	-44%
2681	0.21	0.19	0.01	7%	0.25	-0.04	-17%
5004	0.18	0.5	-0.32	-64%	0.35	-0.17	-48%
2380	0.12	0.38	-0.26	-68%	0.44	-0.32	-72%
2310	0.11	0.36	-0.25	-71%	0.38	-0.28	-73%
355	0.08	0.23	-0.15	-66%	0.04	0.04	93%

Figure Legends

Figure 1. Flowchart of the eICU study population. From the entirety of the patient unit stay IDs in the dataset, the cohort was filtered to those with 'trauma' in the diagnosis string. Encounters with non-acute trauma diagnoses were also excluded. The results were flattened to hospital stay. Hospital stays with missing outcome data were removed from the data set.

Figure 2. Flowchart of the NTDB Study Population. From the entire NTDB database of the 2014 and 2015 encounters, 30 Level II facilities were chosen via stratified random sampling based on hospital size. The cohort was determined by patients who required critical care during the first 24 hours of their hospital stay. Modeling and evaluation were performed at the facility level. Facility-specific NTDB data split with 50% of the data were added to the source data to form the training set, and 50% of the data were used for the validation set.

Figure 3. Modeling Schematic for heterogeneous transfer learning approach. The source and target dataset are combined with three groups of features: 1) features in common between the source and target, 2) features unique to the source, and 3) features unique to the target. The union of the two feature sets becomes the feature set for the training data. Once missing values are imputed and scaled, 50% of the target data is used for model validation. Facility-specific data from the target is used for model testing.

Figure 4. Difference in Magnitude of Correlation Coefficients among Binary Variables, Pre- and Post-Imputation. Correlation coefficients between independent binary variables and the outcome were assessed to determine the impact of imputation on the association. Pre-imputation correlation coefficients evaluated the relationship between independent variables and the outcome in the target data at the facility-level. Post-imputation correlation coefficients evaluated the relationship when the target data was added to the source data and imputation had occurred.

Figure 5. Distribution of F1 Scores by Proportion of Target Data in Training Data. This figure depicts the relationship between the proportion of target data in the training set and the heterogeneous TL derived F1 score. The red markers are those facilities wherein the heterogeneous TL approved outperformed or performed comparably to baselines.

Figure 6. Range of Performance by Modeling Approach. This figure of boxplots describes the range of F1 scores across all facilities by different modeling approaches.

Figures

Figure 1. Flowchart of eICU Study Population

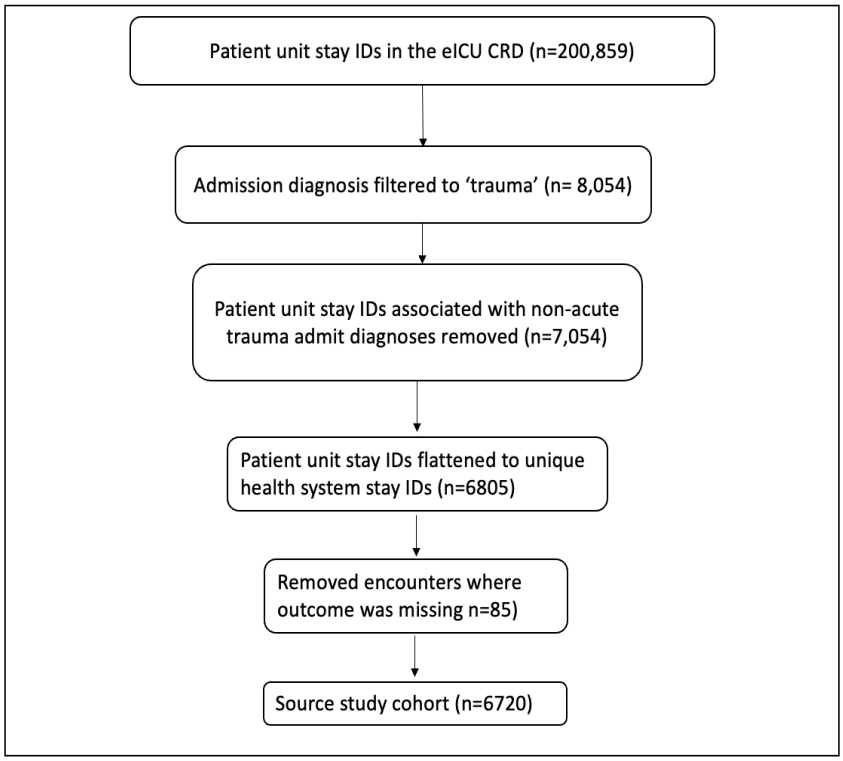


Figure 2. Flowchart of NTDB Study Population

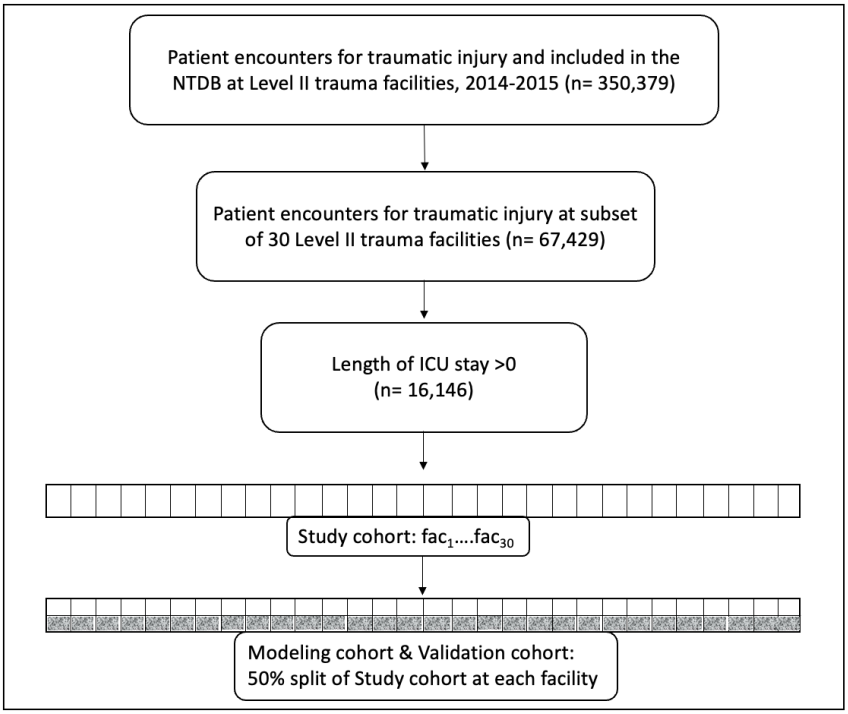


Figure 3. Modeling Schematic of heterogeneous transfer learning approach

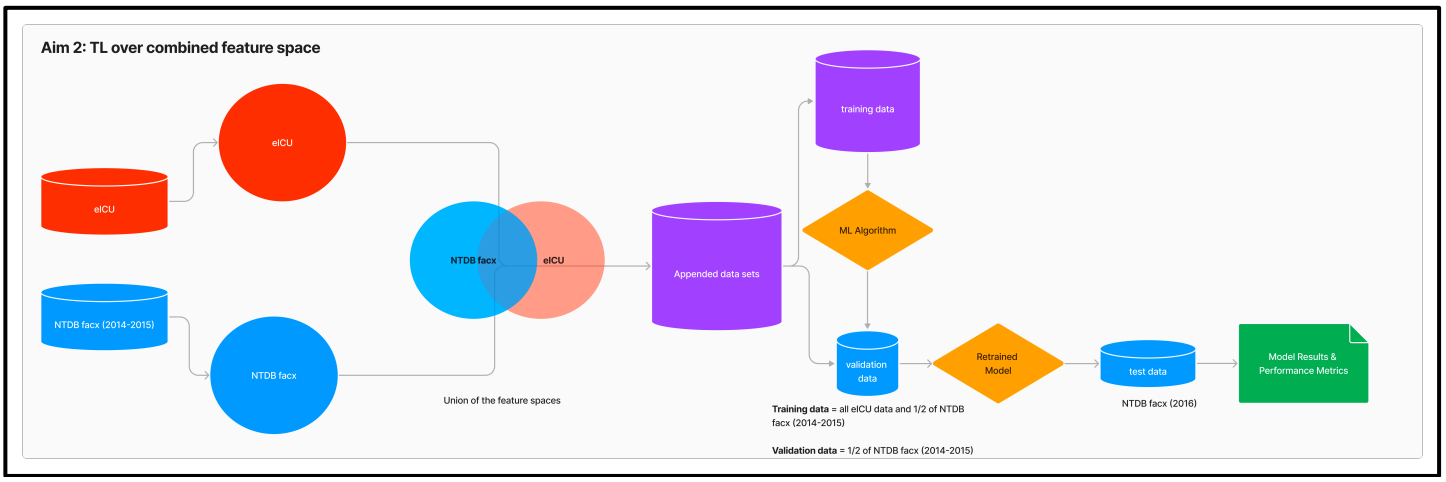


Figure 4. Difference in Magnitude of Correlation Coefficients among Binary Variables, Pre- and Post-Imputation.

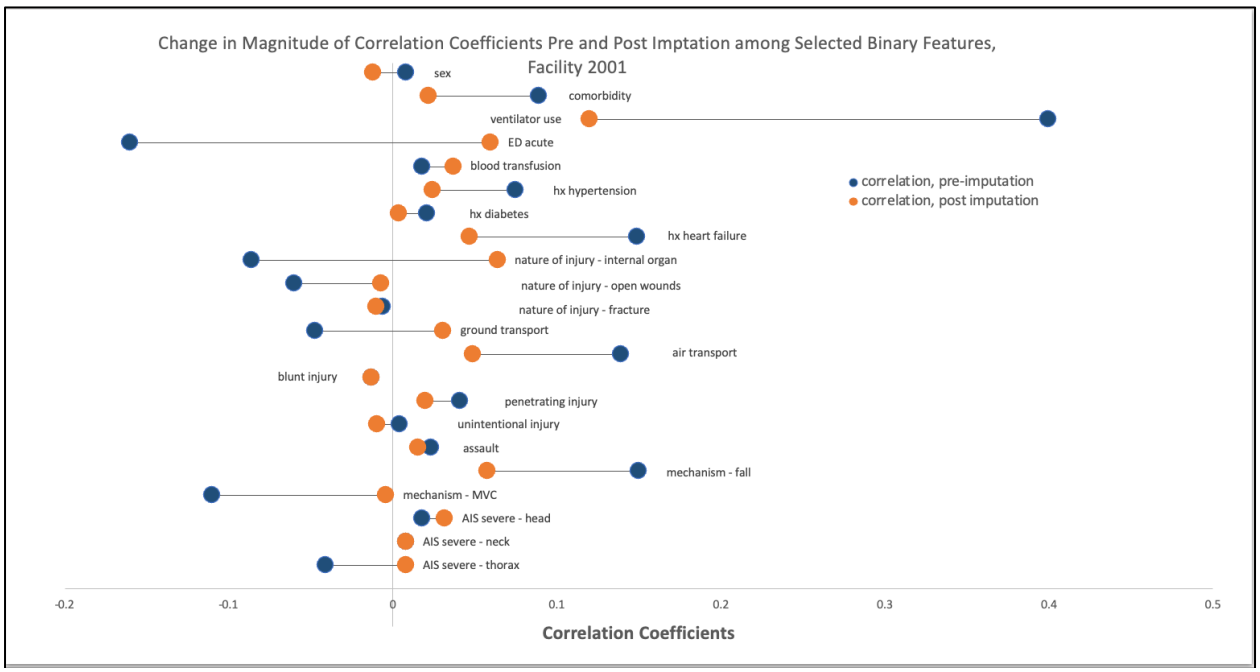


Figure 5. Distribution of F1 Scores by Proportion of Target Data in Training Data

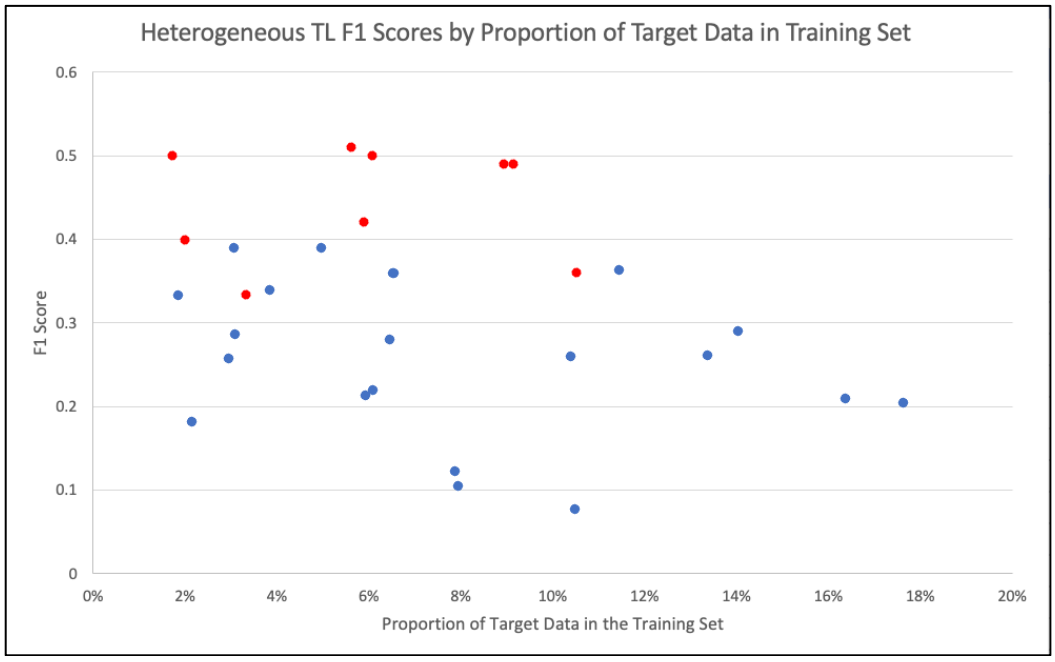
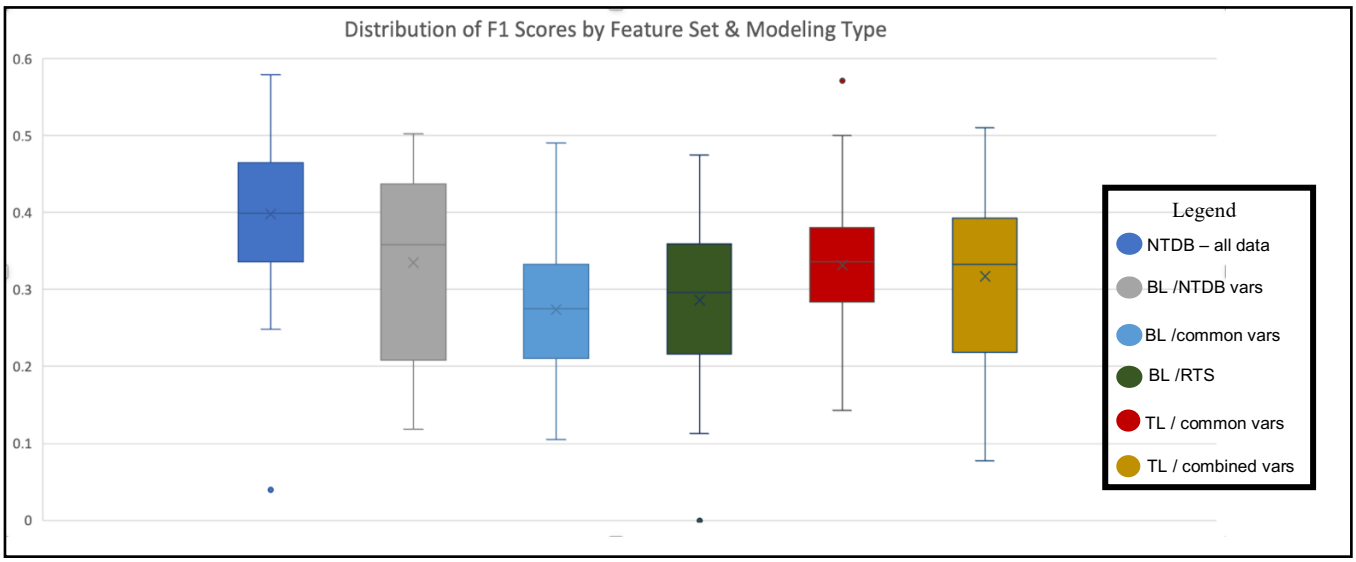


Figure 6. Range of Performance by Modeling Approach (F1 scores)



Supplemental Material

Supplemental Paper 2:

eICU Cohort Determination

As described, the cohort of trauma encounters was determined based on the string 'trauma' in the diagnosis field of the eICU data. However, when evaluating the entirety of the diagnosis string, it became evident that non-traumatic diagnoses remained in the data. The following diagnoses and their associated encounters were excluded from the study cohort:

- Non-traumatic amputation (n=126)
- Facial surgery (n=59)
- Non-traumatic fracture (n=124)
- Non-traumatic hemorrhage (n=338)
- Hip surgery (n=126)
- knee surgery (n=83)
- non-traumatic coma (n=11)
- trauma medical condition (n=83)

Feature Development

The following list describes each feature and the calculations or groupings required for the source and target data to align them in the union of their feature space:

- Outcome:
    - eICU: discharge location of death served as outcome = 1
    - NTDB: discharge disposition of death or hospice served as outcome = 1
- Age: maintained as a continuous variable
    - eICU: In the original data, patients older than 89 years are indicated as age >89 years. This was modified to age = 90 in the dataset
    - NTDB: patient ages in months were all rounded to "0" or "1." Patients aged "≥ 89 years were modified to 90 years.
- Sex
    - eICU: binary variable including male (1) or female (2)
    - NTDB: binary variable including male (1) or female (2)
- Pulse: maintained as a continuous variable
- Systolic Blood Pressure (SBP): maintained as a continuous variable
- Oxygen Saturation (Oxysat): maintained as a continuous variable with a maximum value of 100
    - eICU: component scores and composite obtained from (table / field)
    - NTDB: component scores and composite obtained from (table/ field)
- Body mass index (BMI): once calculated, maintained as a continuous variable
    - eICU: calculated as weight in kg divided by the square of height in meters; outliers were discovered based on these calculations and treated as follows:
        - values less than x:
        - values more than 100: treated as missing (see below)
        - values > 60 and <100, respectively, and modified to 60.
    - NTDB: calculated as weight in kg divided by square of height in meters
        - values less than x:
        - values more than 100: treated as missing (see below)
        - values > 60 and <100, respectively, and modified to 60.
- Glasgow Coma Scale (GCS): treated as an integer
    - eICU: component scores and composite obtained from (table / field)
    - NTDB: component scores and composite obtained from (table/ field)
- blood transfusion: binary indicator defined by transfusion of red blood cells within 24 hours of admission
    - eICU: transfusion defined as (what table / field) with text string including "packed red blood cells", "transfusion of >2 units prbcs", "blood", or "transfusion of 1-2 units of prbcs" (within the prescribed offset of 1440)
    - NTDB: transfusion defined within table 'PCODEDESCR' as "transfusion of paced cells" on day 0 or 1 of admission.
- Ventilator use: binary indicator defined by ventilatory requirement
    - eICU: ventilator use defined using the 'Apache Patient Result' table and defining ventilator use as 'actual vent days' $\geq$ 0 within the prescribed offset of 1440
    - NTDB: ventilator use defined as 'VENTDAYS' with a duration >0

- Specific Comorbidities: Binary features were generated for the following comorbidities based on their shared availability in the source and target datasets:
    - hypertension
    - congestive heart failure (CHF)
    - cerebrovascular accident (CVA)
    - diabetes
    - chronic obstructive pulmonary disease (COPD)
    - myocardial infarction (MI)
    - renal disease
    - cancer
    - eICU: these comorbidities were obtained from the “past history” data file with the following dictionary
        - hypertension:
            - hypertension requiring treatment
        - CHF:
            - CHF - severity unknown
            - CHF - class I
            - CHF - class II
            - CHF - class III
            - CHF - class IV
        - CVA:
            - TIA(s) - within 6 months
            - TIA(s) - within 2 years
            - TIA(s) - within 5 years
            - TIA(s) – remote
            - TIA(s) - date unknown
            - Stroke - within 6 months
            - Stroke - within 2 years
            - Stroke - within 5 years
            - Stroke - remote
            - Stroke - date unknown
        - Diabetes:
            - Medication dependent diabetes
            - Insulin dependent diabetes
            - Non-medication dependent diabetes
        - COPD:
            - COPD - no limitations
            - COPD - moderate
            - COPD - severe
        - MI:
            - MI - within 6 months
            - MI - within 2 years
            - MI - within 5 years
            - MI – remote

        - MI - date unknown
    - Renal Disease:
        - Renal insufficiency - creatine 1-2
        - Renal insufficiency - creatinine 2-3
        - Renal insufficiency - creatinine 3-4
        - Renal insufficiency - creatinine 4-5
        - Renal insufficiency - baseline creatinine unknown
        - Renal failure - not currently dialyzed
        - Renal failure - hemodialysis
        - Renal failure - peritoneal dialysis
    - Cancer:
        - Bladder
        - Brain
        - Breast
        - Cancer
        - CML
        - Colon
        - Esophagus
        - Head and neck
        - Hodgkin's disease
        - Leukemia - other
        - Liver
        - Lung
        - Melanoma
        - Multiple
        - Multiple myeloma
        - Ovary
        - Pancreas - adenocarcinoma
        - Prostate
        - Testes
        - Uterus
- NTDB: these comorbidities were obtained from the "past history" data file with the following dictionary
    - Hypertension:
        - Hypertension requiring medication
    - Diabetes:
        - Diabetes mellitus
    - CHF:
        - Congestive heart failure
    - CVA:
        - CVA/residual neurological deficit
        - Cerebrovascular accident (CVA)
    - COPD:
        - Chronic Obstructive Pulmonary Disease (COPD)

    - MI:
        - History of myocardial infarction
    - Renal Disease:
        - Chronic renal failure
    - Cancer:
        - Disseminated cancer
        - Currently receiving chemotherapy for cancer
- Additionally, a comorbidity indicator variable was generated from the above dictionaries to indicate if a patient had a known history of any of the included comorbidities at the time of injury
- Nature of Injury:
    - eICU: These characteristics were extracted from the final element of the 'diagnosis' string and categorized as follows:
        - Fracture:
            - Bone fracture(s)
            - Fracture of skull
        - Internal organ:
            - Adrenal trauma
            - Bladder trauma
            - Blunt abdominal trauma
            - Bowel trauma
            - Cardiac injury – blunt
            - Diaphragmatic injury
            - Hepatic trauma
            - Intracranial injury
            - Lung trauma
            - Mesenteric trauma
            - Pancreatic trauma
            - Pneumomediastinum
            - Pneumothorax
            - Renal trauma
            - Retroperitoneal trauma
            - Spinal cord injury
            - Splenic trauma
            - Tension pneumothorax
            - Traumatic injury of esophagus
            - Ureteral trauma
        - Open Wounds:
            - Chest wall trauma
            - Penetrating abdominal trauma
            - Pellet injury from shotgun blast
            - Cardiac injury – penetrating
            - Airway trauma

        - Sprain: sprain
        - Dislocation: dislocation
        - Blood vessels:
            - Traumatic injury to major vessel(s)
            - Extremity ischemia
            - Extremity compartment syndrome
            - Blunt vascular injury
            - Penetrating vascular trauma
            - Mesenteric ischemia
        - Amputation: amputation
        - Other: none

Target Specific Features
These additional features were included in the heterogeneous TL models and were derived from the target data.

- Payor type, categorized as follows:
    - Private: private, commercial, Blue Cross / Blue Shield
    - Government: Medicare, Medicaid, Other government
    - Self: self-pay
    - Other: Other, No-Fault Auto, Workers Compensation, Not Billed, Not Known, Not Applicable
- Temperature
- Respiratory Rate
- Supplemental Oxygen: a binary feature generated from 'SUPPOXY' variable. 'Supplemental oxygen' was coded as a 1 while 'no supplemental oxygen,' 'not applicable,' and 'not recorded' were coded as 0.
- Type of Transport:
    - Ground
    - Air
    - Other
- Trauma type: binary features for each of the following types:
    - Blunt
    - Penetrating
    - Burn
    - Other
    - Nan
- Trauma intent: binary features for each of the following intents:
    - Unintentional
    - Assault
    - Self-Inflicted
    - Other
    - NaN
- Trauma mechanism: binary features for each of the following mechanisms:

- Fall
- MVC:
  - MVT Occupant
  - MVT Motorcyclist
  - MVT Pedestrian
  - MVT Pedal cyclist
  - MVT Other
  - MVT Unspecified
- Firearm
- Struck by, against
- Machinery
- Transport:
  - Transport, other
  - Pedal cyclist, other
  - Pedestrian, other
- Other:
  - Unspecified
  - Other specified and classifiable
  - Other specified, not otherwise classifiable
  - Poisoning
  - Natural/environmental, Bites and stings
  - Natural/environmental, Other
  - Suffocation
  - Drowning/submersion
  - Fire/flame
  - Adverse effects, medical care
  - Adverse effects, drugs
  - Hot object/substance
  - Overexertion

Source Specific Features

These additional features were included in the heterogeneous TL models and were derived from the source data.

- Laboratory tests
  - troponin_test (indicator)
  - troponin_positive
  - INR_positive
  - PTT_positive
- Vital signs
  - Highest heart rate
  - Lowest systolic BP
  - Highest respiratory rate
  - Lowest respiratory rate
  - lowest Oxygen saturation

- Physical Exam
    - acute distress
    - irregular rhythm
    - dusky
    - with guarding
    - not reactive
    - agitated
    - obese
    - FFP
- Infusions
    - vasoactive infusions
    - hypertonic infusions
- Treatment
    - aggressive volume repletion
    - blood products
    - inotropes
    - vasopressor
- Home Medications were mapped into 30 therapeutic classes with subject matter expertise. These classes included groups like:
    - anti-hypertensions
    - psychiatric medications
    - analgesic medications
    - GERD medications
    - respiratory medications
    - antidiabetic medications
    - antihyperlipidemia medications
    - anticoagulants

Test Dataset Preparation

Facility-specific NTDB data from 2016 were used as the test set for the TL-based model and all the baseline models. We prepared the data using the following steps:

- Subset patients to those seen in the 30 facilities we used to define the study cohort
- Removed the 45 patient encounters in which 'discharge disposition' was missing
- Subset cohort to those patients with 'ICUDAYS'>0
- Test cohort of 9045 patient encounters across 30 facilities

The feature development proceeded according to the steps outlined in the previous section, with a few exceptions. ICD 10 coding was implemented in October 2015. The NTDB 2016 data includes ICD 9 and 10 codes, however, D-codes and E-codes were less complete for ICD9 coding than for ICD10. Therefore, we used ICD 10 codes and translated them into ICD 9 code groupings to align with the study cohort (2014-2015). For example, 'cut/pierce' was added as an additional mechanism of injury in 2016. We grouped injuries using this mechanism as 'struck.' Similarly, ICD10 procedure codes in the 2016 data had to be mapped to ICD-9 data for the purposes of this analysis. Most categories mapped cleaning from one code grouping to the next, except for 'operations of the ear'. We combined operations of the ear with those of the nose, mouth, and pharynx to align with ICD-10 procedure coding. Code groupings are described in *Supplement: Tables 4* and *5*.

Detailed Methods

Python version 3.8.5 was used for data management, analysis, and modeling. The following packages and libraries were used:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Joblib
- statsmodels.api

Feature Selection

Features with high cardinality were transformed to binary features with one hot encoding. The significance of these features was determined by evaluating the entropy associated with each feature. Features with an entropy of <0.9 were maintained for inclusion in the model development stage. Entropy calculations are shown in Supplement: *Tables 3, 4,* and *5.*

The following additional details accompany the development, training, and testing of the TL model, as well as the 'all features' and 'shared features' baselines:

- LASSO regularization was used to select the most important features in our feature sets
- LASSO used an L1 penalty and liblinear solver
- The selected features were then used in a logistic regression model, which was optimized on the training data using a liblinear solver and evaluating the following values for C: 0.1, 1.0, and 10.0
- The best performing model, according to F1 score, was then evaluated on the validation set with different thresholds for the positive class (0.05, 0.1, 0.15, 0.20, and 0.25)
- The optimal threshold was determined via bootstrapping (=100)
- This best performing model was then used on the test set for final model metrics, with bootstrapping (n=1000) to generate 95% confidence intervals for precision and recall

R, and the icdpicr package was used to map the ICD 9 codes found in the eICU data to the AIS scores.

ChatGPT was used to assist with coding and perform code reviews when errors occurred. Examples of prompts used include “I would like to generate a dot plot to represent many different entities according to a performance metric. I would like to generate multiple subplots, one for each entity. Please provide me with sample code” and “I would like to integrate LASSO regression for feature selection into my modeling code. Please provide me an example of how I would implement LASSO in this chunk of code.”

Supplemental Tables & Figures

Table S1. Facility Characteristics, Training Data (2014-2015)

Facility	Cohort Size	Positive Class	Transfer Rate (%)	ICU Trauma Beds	Trauma Surgeons	Region	Teaching Status
6552	401	22	0.7	26-35	4-5	Northeast	Community
5003	119	15	40.3	11-15	4-5	Midwest	Non-teaching
6547	435	38	11.3	16-25	5-6	South	Non-teaching
2005	787	74	9.4	16-25	4-5	West	Community
2199	678	140	43.4	26-35	4-5	Midwest	Non-teaching
2154	269	32	32.7	26-35	7-8	Midwest	Non-teaching
6131	138	31	47.1	11-15	4-5	West	Community
315	204	37	21.6	26-35	4-5	Midwest	Non-teaching
2404	436	61	29.8	26-35	5-6	Midwest	Non-teaching
358	1439	69	9.85	16-25	4-5	West	Non-teaching
2194	126	14	0	11-15	4-5	Northeast	Community
2705	212	8	5.7	16-25	4-5	South	Community
6362	471	42	15.3	16-25	4-5	Midwest	Community
460	868	59	47.4	>35	4-5	West	Community
2001	351	44	41.3	11-15	4-5	Midwest	Community
6555	232	36	0.9	16-25	4-5	Midwest	Community
138	1097	76	13.9	>35	4-5	West	Community
6268	214	20	15.4	16-25	4-5	West	Non-teaching
2399	660	92	24.1	>35	5-6	Northeast	University
2065	468	63	61.3	>35	>8	Midwest	Community
2416	580	57	26.6	11-15	7-8	West	Community
2680	575	60	33.9	1-10	1-3	Midwest	Community
2254	780	61	15.9	16-25	4-5	West	Non-teaching
2681	463	37	16.6	16-25	5-6	West	Community
6242	424	13	22.4	26-35	5-6	West	Community
6244	1316	137	15.1	16-25	5-6	South	Community
5004	147	13	49.7	11-15	>8	Midwest	University
2380	792	75	33.5	>35	5-6	Northeast	University
2310	421	23	12.4	11-15	4-5	West	Community
355	1037	82	6.8	>35	7-8	West	Community

Table S2. Characteristics of Treated Patients by Facility, 2014-2015

	Patient Age (med)	Youngest Age Treated	Penetrating Trauma (%)	ISS (med)
6552	57	16	4.7	16
5003	37	<1	9.2	17
6547	54	15	5.3	17
2005	45	15	11.6	16
2199	52	<1	10.2	17
2154	57	15	5.6	17
6131	40	13	6.5	22
315	58	16	4.9	13
2404	54	<1	9.2	16
358	52	9	3.9	14
2194	62	18	1.6	17
2705	65	15	4.2	13
6362	56	<1	9.8	17
460	47	<1	5.8	16
2001	63	16	4.6	16
6555	69	14	4.7	17
138	52	14	5.5	17
6268	56	15	5.1	17
2399	62	16	7.6	17
2065	51	<1	5.6	17
2416	56	14	7.6	14
2680	57	<1	4.3	10
2254	55	15	5.6	17
2681	61	16	2.8	16
6242	39	12	13.2	17
6244	50	<1	8.0	14
5004	48	8	5.4	20
2380	54	<1	5.9	17
2310	42	10	11.6	13
355	51	<1	2.1	16

Table S3. Entropy Evaluation for Source Specific Features

Source	Feature	Definition	Prevalence	Entropy	In Model
Lab	trop_bin	Troponin-I lab test occurred	19.29%	.954	No
	trop_pos	Troponin-I > 0.4 ng/ml	2.44%	.999	No
	INR_pos	INR > 1.5	10.43%	.75	Yes
	PTT_pos	PTT > 40 sec	4.6%	.61	Yes
Physical Exam	HR_highest	Continuous value	67%	--	Yes
	Syst BP lowest	Continuous value	73%	--	Yes
	Resp_highest	Continuous value	63%	--	Yes
	Resp_lowest	Continuous value	63%	--	Yes
	Sats_lowest	Continuous value	66%	--	Yes
	acute_distress	Text string	0.87%	.59	Yes
	irregular_rhythm	Text string	1.0%	na	No
	dusky	Text string	0.01%	na	No
	with_guarding	Text string	.09%	na	No
	not_reactive	Text string	0.58%	na	No
	agitated	Text string	1.6%	.98	No
	obese	Text string	1.1%	.97	No
	FFP	Text string	1.7%	.65	Yes
Infusion Drugs	Vasoactive	Any of the following: Norepinephrine, vasopressin, phenylephrine, epinephrine, dobutamine, dopamine, milrinone	6.6%	.58	Yes
	Hypertonic	Text string with ‘3%’	0.34%	.995	No
Treatment	Aggress_volume	Text string	0.37%	.995	No
	Blood_products	Text string	10.6%	.87	Yes
	Inotropes	Text string	1.3%	.61	Yes
	Vasopressor	Text string	6.8%	.62	Yes
Admit Meds	Anti_htn	Therapeutic med groups	7.7%	.99	No
	Psych	Therapeutic med groups	3.6%	.99	No
	Analgesic	Therapeutic med groups	3.7%	.99	No
	GERD	Therapeutic med groups	3%	.95	No
	Respiratory	Therapeutic med groups	2.1%	.98	No
	Antidiabetic	Therapeutic med groups	0.94%	.96	No
	Antihyperlipidemia	Therapeutic med groups	4.3%	.99	No
	Anticoagulant	Therapeutic med groups	3.2%	.95	No

Table S4. Entropy Evaluation of Target Specific Features, E-code Groupings (ICD-9 CM)

E code group	Codes included	Prevalence	Entropy	Included?
External cause status	0, 0.99	--	--	No
Activity	1, 30.99	--	--	No
Railway	800, 808.99	11	--	No
MVA	810, 819.99	5118	.99	No
Motor vehicle non-traffic	820, 825.99	558	.93	No
Other Road Vehicle	826, 829.99	467	.73	Yes
Water Trans Accidents	830, 838.99	49	--	No
Air & Space Transport	840, 845.99	20	--	No
Vehicle accidents NOS	846, 848.99	5	--	No
Place of occurrence	849, 849.99	--	--	No
Accidental poisoning	850, 858.99	2	--	No
Accidental falls	880, 888.99	6681	.98	No
Accidents / fire or flames	890, 899.99	20	--	No
Accidents / environment	900, 909.99	57	--	No
Submersion or suffocation	910, 915.99	22	--	No
Other accidents	916, 928.99	600	.95	No
Suicide / self inflicted	950, 959.99	360	.83	Yes
Homicide / assault	960, 969.99	1375	.38	Yes

Table S5. Entropy Evaluation of Target Specific Features, Procedure Code Groupings (ICD-9 CM)

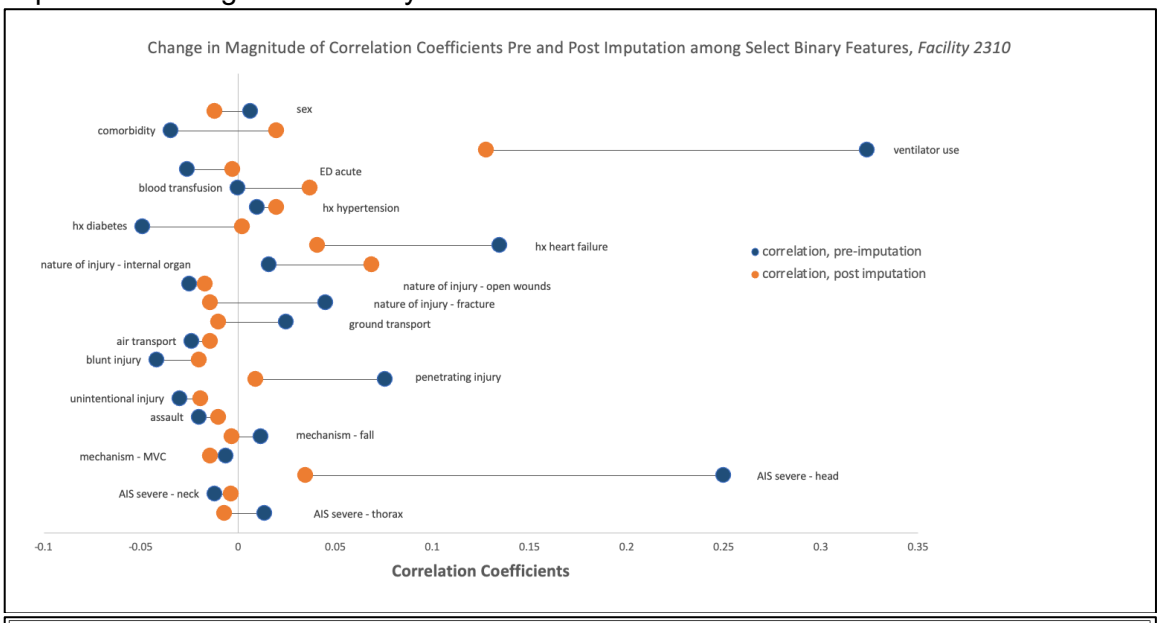
P code group	Codes included	Prevalence	Entropy	Included?
Procs NOS	0.01, 0.96	1.1%	.97	No
Operations / nervous system	1.01, 5.9	12.8%	.89	Yes
Operations / endocrine system	6.01, 7.99	0.17%	0	No
Operations / eye	8.01, 16.99	2.0%	.89	Yes
Misc procedures	17.11, 17.81	0.1%	0	No
Operations / ear*	18.01, 20.99	0.62%	.98	No
Operations / nose, mouth, pharynx	21.00, 29.99	2.8%	.89	Yes
Operations / respiratory system	30.01, 34.99	11.7%	.97	No
Operations / cardiovascular system	35.00, 39.99	20.0%	.91	No
Operations / heme and lymph	40.00, 41.99	1.7%	.97	No
Operations / digestive system	42.01, 54.99	7.8%	.99	No
Operations / urinary system	55.01, 59.99	14.6%	.98	No
Operations / male GU	60.00, 64.99	0.1%	.99	No
Operations / female GU	65.01, 71.9	0.1%	0	No
OB procedures	72.0, 75.99	0.15	.93	No
Musculoskeletal procedures	76.01, 84.99	19.5%	.94	No
Operations / integumentary system	85.0, 86.99	19.7%	.97	No
Misc	87.01, 99.99	76.1%	.999	No

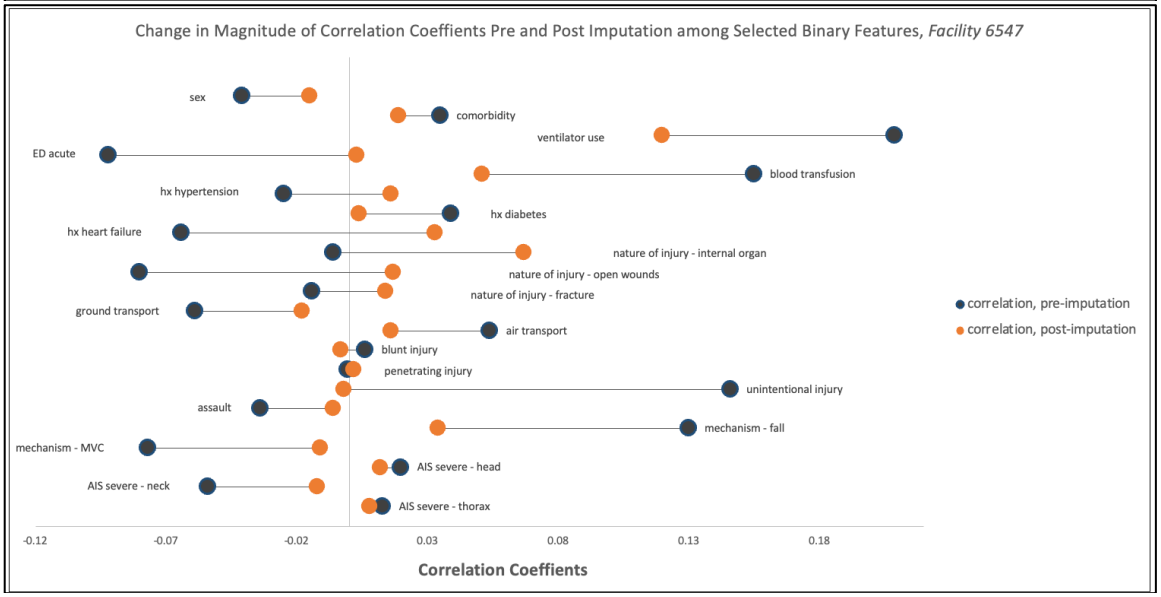
Supplemental Table S6.

Degree of Misalignment (number of features)	Number of Facilities (n)	Mean F1 (Heterogeneous TL)
0-2	5	0.35
3	5	0.34
4	8	0.26
5-6	6	0.36
7+	6	0.33

* Among the shared feature set, conditional probabilities were compared between the source data and each facility's target data set. The conditional probability distributions were compared, and their difference was evaluated for significance with Fisher's Exact Test. A misaligned feature was counted when the difference between the probability distributions of the target and source was significant at the .05 level.

Supplemental Figure 1. Differences in Magnitude of Correlation Coefficients Pre and Post Imputation among Select Binary Features





Conclusion

Our results suggest a more nuanced view of the performance effects of TL from those of other studies that demonstrate its performance benefits. When developing a risk model to predict *Clostridium difficile (c. difficile)* infection risk at admission across three hospitals, Wiens *et al.* determined that a TL approach resulted in the best-performing models assessed via AUROC and AUPR scores.[1] In another study, Desautels et al. used the MIMIC III dataset, which includes data from more than 50,000 intensive care unit stays, to augment patient mortality prediction among encounters from the University of California San Francisco Medical Center.[2] The authors demonstrated that their homogeneous TL model outperformed the commonly used inpatient mortality prediction tool, the Modified Early Warning Score. Finally, using the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) and electronic health record data from a single hospital, Lee et al. used TL to develop a model to predict mortality following surgery, an approach that resulted in improved performance over comparative models.[3] In each preceding example, a TL approach uses a larger record set to augment the training data available to predict local patterns in patient outcomes. However, there are important differences in our approach that may explain why our findings differ.

1. Effect of Dataset Provenance

Based on the published literature on TL, researchers caution that the success of this approach is not guaranteed, and that negative transfer can occur. The risk of negative transfer, in which the predictive performance is reduced from the baseline, correlates with the degree of dataset misalignment.[4] One author advises that the source and target task should be of related domains to reduce the likelihood of negative transfer.[5] Other papers comment on the importance of conditional and marginal probability distributions in reducing the likelihood of negative transfer.[5,6] Based on our experience, it seems that another element to consider should be the provenance of the dataset. That is how was it developed, and what was its original intent. Consider, our target data, the NTDB. The NTDB is a robust trauma registry developed for the evaluation and research of trauma systems that has become increasingly standardized over time.[7] Trauma centers have dedicated personnel tasked with extracting data elements from patient charts for the NTDB, thereby increasing the fidelity of data capture in a way not replicated in electronic health record (EHR) extracts.[8] As such, the NTDB explicitly extracts many trauma-related variables that are key to trauma outcome prediction, which may not be

readily available in an EHR database. The eICU dataset, our source data, draws data from a heterogeneous group of hospitals that, while using the same data capture system, have no standardization or requirements for data collection and reporting beyond that required for patient care.[9]

Comparing this to other published research, in the study by Wiens et al. the authors used data extracted from the EHR of three related hospitals, albeit with some differences in feature sets[1]. The three data extracts were similar in the time period of care they reflected, their geography, and their patient inclusion criteria, as all inpatient visits were eligible for inclusion. Desautels et al. used MIMIC-III, an EHR extract, as the source data, and a UCSF EHR extract as the target.[2] Based on our experience, the provenance of the source and target datasets is an important factor to consider and fully understand when beginning experimentations with TL.

2. Cohort Selectivity

To increase the prevalence of the positive class in our data, we limited our dataset to patients requiring intensive care. By selecting this inclusion criteria, we may have increased the difficulty of our prediction. One of the first things young doctors learn is how to triage patients into 'sick' and 'not sick,' as the acuity of care, required resources, and treatments tend to diverge based on this distinction. However, discerning who might survive and who might die is a more challenging task among the already sick and severely injured. In trauma care, patients sometimes 'declare themselves' by presenting a physiological finding or a specific response associated with mortality. Although we found some of these signifiers in the source data (e.g., non-reactive pupils, acute distress), their prevalence was very low. The other work we discussed, that of Wiens, Desautels, and Lee, did not apply such inclusion requirements and instead included all inpatient admissions, all adult admissions, all patients undergoing surgery into their modeling cohorts, respectively.[2,3,1] Because these evaluations used TL across a more general cross-section of patients, the pattern of features associated with their outcome may be more readily discernible.

3. Representation of the Target Data in the Training Data

In TL, the target data is of smaller size than the source.[10] With a small enough target to source ratio, the ML algorithm may ignore the target data when assessing the patterns between predictors and the outcome. In the 30 facilities we modeled, the prevalence of the target data in the training data ranged from 2% to 18% (Figure 4). There are several approaches to amplify

the prevalence of the target dataset in the training data. One approach is to select only the rows from the source data that are most like the target, as defined by various distance metrics, such as cosine similarity or Kullback-Leibler divergence.[11] We considered these approaches but did not apply them because we hypothesized that the fullest extent and variability among the source data would best serve our ML algorithm while avoiding overfitting of the target data. Another approach, similar to oversampling the positive class in cases of class imbalance,[12] is to oversample the target data. This approach can also lead to overfitting. A third approach is to vary the weight of the source and target data, an approach used by Desautels *et al*.[2,6] The authors incorporated the weight of the target data as an experimental variable in their work. They demonstrated most consistent performance when the target data instances were weighted five times that of the source. While not altering the weights of the datasets, Lee et al. used a two-staged development approach beginning with a cost-sensitive support vector machine to train a model on the source data.[3] Then they trained a model on the target data while using the hyperparameters and objective function from the source. This approach allows the target model to learn from the target specific data distributions while transferring the learned hyperparameters from the source data. These approaches may have served to amplify the signal in the target data and to improve learning, as seen in the superior performance of the TL approach.

4. Effect of Imputation

Among the 80 features in our model, the source and target datasets have only 29 features in common. Because most features exist in the source or target, we used feature creation imputation to impute the unknown values related to those features.[13] We evaluated all the data tables in the source and target data to expand the features of our prediction models. Although we considered many features across both datasets, most additional features were unique to the NTDB, such as injury severity scores and e-codes. Once combined with the source data for model training, the values of these features were known in 2-18% of the instances, depending on the relative size of the NTDB data to the eICU. This reduced portion of known data served as the substrate for imputing the remainder of the values. It is very challenging to meaningfully impute values in this situation as the small proportion of known values must either be replicated to match the size of the source data to then use a model like k nearest neighbors, or a single value imputation approach can be used.

In their work, Wiens et al. used target specific features in their heterogeneous TL model, but they omit unique features from the source. Instead of imputing missing values for the target specific features, the authors 'zero-padded' the variables, which were all binary. The authors remark that 'target-specific features are important' based on changes in model performance when those features are omitted.[1] In their work on imputation in TL, Wu et al. first define the two-step imputation approach we used in our work (cross feature imputation followed by feature creation imputation). However, instead of directly imputing values in the feature creation step, they implemented a nearest pairing method to match target instances with source instances before applying a canonical correlation analysis (CCA) function to these matched pairs. The CCA approach resulted in a common latent space which was then used by their algorithm.[13] Target specific features are a key element of TL and the desire to include them is one of the motivations to undertake TL. And yet, imputing the unknown values of these features to the source instances in a way that can be meaningful to the model without overfitting is very challenging. The handling of features unique to the target should be thoroughly discussed among researchers before embarking on this work.

5. Standardization of Care Processes

One of the hypotheses of this study was that facility-specific processes for care after injury may supersede more general care processes that are important for trauma prediction models. However, injury care may be a poor choice for demonstrating TL. The inclusion criteria for facilities in our study included ACS Level II certification, a rigorous requirement.[14] In addition, trauma care itself is standardized by widely published guidelines and algorithms.[7] The results of this standardization are well published and have led to superior care being provided to trauma patients in ACS-verified centers, particularly ACS level II centers when compared to their state-verified counterparts.[15] The local effects of facility-specific processes may not be relevant when studying traumatic injuries. We can juxtapose the standardization of trauma care with that of *C. difficile* prediction among community members (at time of admission), inpatient mortality among a heterogeneous group of medical patients, and mortality prediction across a wide range of surgical patients. In short, the care pathways used among trauma patients may not reflect the heterogeneity of the care pathways and potential outcomes we had hypothesized.

This work represents a robust approach to evaluating the use of TL in a variety of contexts to predict outcomes following traumatic injury. The variety of techniques, the practicality of the

approach, and the heterogeneity of facilities represented in the data provide ample perspectives on the efficacy of TL in this context. Through this work, we uncovered several insights to guide future efforts in TL research. We suggest that these insights be used to further develop clear criteria to guide the use of TL.

References

1. Wiens J, Guttag J, Horvitz E. A study in Transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*. 2014;21(4):699-706. doi:10.1136/amiajnl-2013-002162

2. Desautels T, Calvert J, Hoffman J, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical Informatics Insights*. 2017;9:117822261771299. doi:10.1177/1178222617712994

3. Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. *2012 IEEE 12th International Conference on Data Mining Workshops*. Published online 2012. doi:10.1109/icdmw.2012.93

4. Wang Z, Dai Z, Poczos B, Carbonell J. Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2019. doi:10.1109/cvpr.2019.01155

5. Weiss K, Khoshgoftaar TM, Wang D. A survey of Transfer Learning. *Journal of Big Data*. 2016;3(1). doi:10.1186/s40537-016-0043-6

6. Curth A, Thoral P, van den Wildenberg W, et al. Transferring clinical prediction models across hospitals and Electronic Health Record Systems. *Machine Learning and Knowledge Discovery in Databases*. Published online 2020:605-621. doi:10.1007/978-3-030-43823-4_48

7. Hashmi ZG, Kaji AH, Nathens AB. Practical guide to surgical data sets: National Trauma Data Bank (NTDB). *JAMA Surgery*. 2018;153(9):852. doi:10.1001/jamasurg.2018.0483

8. Committee on Trauma American College of Surgeons. Resources for Optimal Care of the Injured Patient: American College of Surgeons, 2014. American College of Surgeons website. Available at: https://www.facs.org/-/media/files/quality-programs/trauma/vrc-resources/resources-for-optimal-care.ashx

9. 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). doi:10.1038/sdata.2018.178

10. Pan SJ, Yang Q. A survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2010;22(10):1345-1359. doi:10.1109/tkde.2009.191

11. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951 Mar 1;22(1):79-86

12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-57.

13. Wu X, Khorshidi HA, Aickelin U, Edib Z, Peate M. Transfer learning to enhance amenorrhea status prediction in cancer and fertility data with missing values. Artificial Intelligence: Applications in Healthcare Delivery. 2020 Dec 2:233.

14. Brown JB, Watson GA, Forsythe RM, er al. American College of Surgeons trauma center verification versus state designation: are Level II centers slipping through the cracks?. The journal of trauma and acute care surgery. 2013 Jul;75(1):44.

15. Sakran JV, Jehan F, Joseph B. Trauma systems: standardization and regionalization of care improve quality of care. Current Trauma Reports. 2018 Mar;4:39-47.