Original Paper

Development and Evaluation of Embeddings for Representing Injuries in Patients With Traumatic Injuries: Retrospective Cohort Study

Kelvin Szolnoky[1], MD; Jonatan Attergrim[2,3], MD; Awais Ashfaq[4,5], PhD; Henrik Linusson[6], PhD; Martin Gerdin Wärnberg[2,3], MD, PhD; Johanna Berg[2,7], MD, PhD

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Stockholm, Sweden

[2]Department of Global Public Health, Karolinska Institutet, Stockholm, Stockholm, Sweden

[3]Perioperative Medicine and Intensive Care, Karolinska University Hospital, Stockholm, Stockholm, Sweden

[4]Center for Research and Innovation, Region Halland, Halmstad, Halland, Sweden

[5]School of ITE, Halmstad University, Halmstad, Halland, Sweden

[6]Ekkono Solutions AB, Varberg, Halland, Sweden

[7]Department of Emergency and Internal Medicine, Skåne University Hospital, Malmö, Skåne, Sweden

Corresponding Author:
Kelvin Szolnoky, MD
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Nobels väg 12a
Stockholm, Stockholm, 171 77
Sweden
Phone: 46 704168708
Email: kelvin.szolnoky@ki.se

Abstract

Background: Patients with traumatic injuries present with heterogeneous injury patterns that are difficult to represent in statistical models. Common approaches either retain individual injury codes in high-dimensional sparse representations, such as one-hot encoding, or aggregate injuries into summary scores, reducing dimensionality at the cost of patient-specific detail.

Objective: This study aimed to develop low-dimensional data-driven ICD-10 embeddings for traumatic injuries and compare their downstream discriminatory performance with one-hot encoding of the same injury codes.

Methods: We conducted a retrospective study using National Trauma Data Bank (NTDB) records from 2018 and 2019. *International Classification of Diseases, Tenth Revision, Clinical Modification* diagnoses were mapped to World Health Organization ICD-10 codes, retaining traumatic-injury codes S00-S99 and T15-T34. A denoising autoencoder was developed on 2018 data and temporally evaluated on 2019 data. Representations with 2, 4, 8, 16, 32, and 64 dimensions were compared with one-hot encoding across 16 downstream tasks using logistic regression and LightGBM. Performance was measured by the area under the receiver operating characteristic curve (AUC) for binary tasks and macro one-vs-rest AUC for multiclass tasks. Models were trained and evaluated within individual hospitals and separately on the full NTDB. Hospitals were stratified into volume categories, from small (≤500) to major (2500 patients/year). CIs were obtained from 1000 nonparametric bootstrap samples.

Results: The development cohort included 1,041,021 eligible patients from 2018, and the temporal evaluation cohort included 1,094,933 patients from 2019. At the hospital level, embeddings outperformed one-hot encoding. With LightGBM at 16 dimensions, the mean paired AUC difference across tasks was 0.054 (95% CI 0.048-0.060) in small hospitals (n=120) vs 0.023 (95% CI 0.020-0.025) in major hospitals (n=99). With logistic regression, differences were 0.056 (95% CI 0.050-0.063) and 0.041 (95% CI 0.039-0.043), respectively. In the full NTDB, 803-indicator one-hot encoding outperformed embeddings. Mean differences were –0.004 (95% CI –0.004 to –0.003) with LightGBM and –0.030 (95% CI –0.030 to –0.029) with logistic regression. At 64 dimensions, these differences narrowed to –0.001 and –0.016, respectively.

Conclusions: Embeddings with as few as 8 dimensions achieved predictive performance comparable to or better than one-hot encoding, with up to a 100-fold reduction in dimensionality. Compared to one-hot encoding, embeddings improved performance in models trained at individual hospitals, particularly lower-volume hospitals. These gains may reflect reduced dimensionality,

XSL•FO
RenderX

parameter sharing across codes, and representation of codes absent from hospital-specific training data, enabled by embeddings learned from the larger development cohort. In models trained on the full registry, one-hot encoding achieved higher mean AUC with both estimators, although the differences were minor (<0.005 AUC difference) for LightGBM with 16- to 64-dimensional embeddings. The embeddings and software are openly available.

(JMIR Med Inform 2026;14:e91298) doi: 10.2196/91298

KEYWORDS

clinical informatics; ICD-10; machine learning; multiple trauma; representation learning; wounds and injuries

Introduction

Trauma comprises physical injury and the body's response [1]. Patients with traumatic injuries have diverse injuries, mechanisms, and severities [2,3]. This heterogeneity complicates the representation of injuries in a form suitable for statistical and machine-learning models [4]. Injuries are commonly coded using the Abbreviated Injury Scale or the *ICD* (*International Classification of Diseases*) [5]. Analytical pipelines often represent these individual codes using binary indicators [6]. Although this retains code-level information, it produces a sparse feature space whose dimensionality increases with the number of distinct codes.

To reduce the number of features, injury patterns are often summarized by mechanism, regional severity, or the Injury Severity Score (ISS) [3,4]. Such summaries necessarily omit some patient-specific detail, and clinically different injury patterns may produce the same score [7,8]. Current approaches therefore involve a trade-off between retaining individual codes in a high-dimensional representation and reducing injuries to a less detailed summary.

Embeddings represent entities, here injury diagnosis codes, as numerical vectors, with entities appearing in similar contexts assigned similar vectors. They can retain distinctions between individual codes while using fewer features than binary encoding. Embeddings of coded patient histories have improved prediction of future diagnoses and readmission compared with models using the coded features directly [9,10]. Models using embeddings have also achieved higher predictive performance than models using one-hot encoding of the same codes and shown improved external generalization [11-13]. Recent models have been pretrained on billions of clinical events, incorporating imaging and clinical text alongside coded data, and can generate patient embeddings for prediction across a range of clinical outcomes [14,15]. Such embeddings incorporate information on past medical history, comorbidities, and acute illness. Embeddings have a greater improvement in predictive performance when task-specific training data is limited [16,17].

Patient embeddings generated from longitudinal medical records capture important clinical information, but are not specifically designed to represent traumatic injuries. These models learn from medical records covering a wide range of conditions, often using the sequence of clinical events to construct patient representations. Injury codes form a small yet heterogeneous subset of these records. Such detailed clinical histories may not be available in trauma registries, which generally collect a defined set of variables relating to an injury admission [18]. These embeddings also contain several hundred dimensions per patient, which may limit their use in statistical models fitted to small datasets [9,15]. To the best of our knowledge, no embedding model has been developed specifically for traumatic injuries.

We therefore aimed to develop compact embeddings of patients' *ICD-10* (*International Classification of Diseases, Tenth Edition*) injury codes that preserve information relevant to prediction. We evaluated these embeddings against one-hot encoding by comparing discrimination across a range of prediction tasks.

Methods

Study Design

We conducted a retrospective study in 2 phases: embedding model development and embedding model evaluation. To assess whether the embeddings preserved information contained in the *ICD-10* codes, we compared embeddings of different dimensions with one-hot encoding using linear (logistic regression) and nonlinear (LightGBM) prediction models across tasks relevant to trauma care. Predictive performance in these tasks served as a proxy for information preservation, based on the assumption that accurate prediction across multiple tasks requires the representation to retain the relevant content of the original coded injuries.

Data and Participants

We used the National Trauma Data Bank (NTDB), a trauma registry maintained by the American College of Surgeons to which participating hospitals in the United States submit data [19]. The NTDB includes patients with a traumatic-injury *ICD-10* diagnosis code who are admitted, transferred, or die. Patients prompting trauma team activation are also eligible irrespective of the discharge diagnoses. Contributing hospitals span level I-V trauma centers and undesignated centers; thus, the registry covers the range of settings in which trauma care is delivered. Because the NTDB does not contain identifiers that allow records to be linked to individual patients, reported patient numbers refer to registry records and may include multiple records from the same individual.

To ensure adequate representation of injuries while avoiding potential confounding effects from the COVID-19 pandemic, we included NTDB data from January 1, 2018, to December 31, 2019. The 2018 patient records were used for embedding model development, hyperparameter selection, and downstream model training. The 2019 patient records were held aside for temporal evaluation. The evaluation dataset was only used to

XSL•FO
RenderX

test the final selected embedding model, ensuring no data leakage between development and evaluation phases.

We included patients with at least one traumatic-injury *ICD* code. We did not restrict inclusion for the embedding model training by patient characteristics, Abbreviated Injury Scale codes, injury severity, or outcome availability. Records with a missing outcome were excluded only from the corresponding downstream task.

Data Preparation

The NTDB uses *ICD-10-CM* (*International Classification of Diseases, Tenth Revision, Clinical Modification*), a modification of *ICD-10*, with a higher level of granularity. To make the embeddings usable across settings, we converted *ICD-10-CM* codes to the World Health Organization (WHO) *ICD-10* codes. We mapped each *ICD-10-CM* code to the longest matching prefix in the WHO *ICD-10* catalog. For example, *ICD-10-CM* code S21.122A ("Laceration with foreign body of the left front wall of the thorax without penetration into the thoracic cavity, initial encounter") was mapped to the international *ICD-10* code S21.1 ("Open wound of the front wall of the thorax"). We only retained diagnosis codes indicating an initial encounter and excluded those indicating subsequent encounters or sequelae of earlier injuries. The mapping consequently removed *ICD-10-CM* detail, including laterality. We retained traumatic injuries in S00-S99 and T15-T34. Other diagnoses, comorbidities, external-cause codes, and procedures were not included in either injury representation method as they do not represent the current injury.

Embedding Model Development

Architecture

We developed a denoising autoencoder to reconstruct each patient's one-hot encoded representation of *ICD-10* injury codes [20]. The encoder output formed the patient embedding, with its dimension determined by the width of the bottleneck layer. We corrupted the input by randomly removing recorded injury codes and adding codes that were not recorded for that patient. The reconstruction target remained the original, uncorrupted representation.

The encoder comprised fully connected hidden layers with layer normalization, rectified linear unit activation, and dropout, followed by a linear bottleneck. A mirrored decoder returned one logit for each injury code included in the model. We trained the network using the AdamW optimizer to minimize binary cross-entropy loss. A tuned positive-class weight adjusted the contribution of present injury codes to the loss to account for their relative sparsity.

Training and Hyperparameter Selection

To preserve the integrity of our 2019 evaluation data for final evaluation, the 2018 hospitals were assigned in a 60:20:20 ratio to training, validation, and test partitions for embedding model development. During embedding model hyperparameter selection, the training partition was used for fitting the model, the validation partition for early stopping, and the test partition for the downstream predictive performance evaluation. The final autoencoder was fitted on the 2018 data.

We independently tuned the embedding model at dimensions of 2, 4, 8, 16, 32, and 64, using 50 trials per dimension drawn by a tree-structured Parzen estimator sampler [21]. The search covered the network architecture, including hidden width and depth, the optimizer, including learning rate and weight decay, and the corruption and loss weighting. The selection objective combined mean area under the receiver operating characteristic curve (AUC) across the 16 tasks in pooled and hospital-specific LightGBM models in the 2018 test partition. Search spaces and selected values are reported in Table S1 in Multimedia Appendix 1.

Embedding Model Evaluation

Baseline Comparison

We used one-hot encoding of the *ICD-10* codes as a baseline representation method for comparison with the embeddings. The one-hot encoding created a matrix where each row represented a patient and each column represented an *ICD-10* code. During the evaluation phase, the embedding model was never retrained or fine-tuned, and its parameters remained fixed after fitting on the 2018 data.

Downstream Prediction Tasks

To evaluate the embeddings for injury representation, we evaluated their predictive performance across several downstream tasks. We compared models using embeddings as features against models using one-hot encoding. For each patient, we constructed both representations from the same recorded *ICD-10* injury codes and used them as the sole input features to logistic regression and LightGBM classification models [22,23]. We included 16 downstream classification tasks: in-hospital mortality (binary), emergency department disposition (multiclass), hospital disposition (multiclass), hemorrhage-control surgery (binary), blood transfusion within 4 hours (binary), withdrawal of life-sustaining treatment (binary), cerebral monitoring (binary), a Glasgow Coma Scale (GCS) score of 8 or less (binary), intensive care unit length of stay (binned multiclass), ISS (binned multiclass), injury mechanism (multiclass), injury intent (multiclass), trauma type (multiclass), performed computed tomography (binary), any operative procedure (binary), and endotracheal airway insertion (binary). Definitions are reported in Table S2 in Multimedia Appendix 1.

The performance on these downstream tasks served only as an internal benchmark against one-hot encoding, rather than benchmarking against other prediction models for these tasks. Only the embeddings and one-hot encoded data were used in logistic regression and LightGBM classification models. No age, sex, GCS score, vital sign, comorbidity, procedure, or other clinical variable was included alongside either injury representation. Consequently, missing values in these descriptive variables did not require complete-case selection or imputation. The aim was to compare the information available from the 2 injury representations under the same downstream models, rather than to construct optimized clinical prediction models.

XSL•FO
RenderX

Hospital-Specific and Full NTDB Evaluation

To assess embedding performance across different training sample sizes, we fitted and evaluated downstream prediction models separately for each hospital. One downstream prediction model was fitted per hospital, task, and representation method. The prediction models were fitted on each hospital's 2018 data and evaluated on the same hospital's 2019 data. Hospitals that did not report during both 2018 and 2019 were excluded from this analysis. For hospital-specific models, each one-hot representation included only codes observed in that hospital's 2018 data. Because of this inherent limitation of one-hot encoding, codes not observed during training were excluded from the 2019 one-hot representation and contributed no information to the model.

The prediction models were also evaluated at the national scale using the full NTDB cohort for 2018 and 2019. One downstream model per task and representation was fitted on all eligible 2018 patients with a nonmissing task label and evaluated on the corresponding 2019 patients.

Performance Metrics

Discrimination was quantified by the AUC for binary tasks, and by the macro one-vs-rest AUC for multiclass tasks. For the full NTDB evaluation, pointwise 95% basic bootstrap CIs were obtained from 1000 nonparametric bootstrap samples, with individual patients as the resampling unit. Hospital-level 95% CIs used the hospital as the resampling unit, with 1000 bias-corrected and accelerated bootstrap samples. We classified each task according to whether the unrounded 95% CI of its paired AUC difference lay entirely above zero (favoring embeddings), entirely below zero (favoring one-hot encoding), or included zero (inconclusive). These outcomes were counted separately within each embedding dimension, downstream model, and evaluation group. No minimum effect size was required. Hospital-specific intervals resampled hospitals, whereas full NTDB intervals resampled patients. These counts therefore reflect both effect magnitude and precision within each evaluation group and should not be used to infer differences in benefit between groups.

Subgroup Analyses

To evaluate the generalizability of the embeddings across different hospital sizes, we stratified results of the hospital-specific evaluation by hospital size. As the NTDB does not include designated trauma center levels, we used annual trauma admission volumes as a proxy for hospital size. We analyzed the distribution of hospitals in the NTDB (Figure S1 in Multimedia Appendix 1), which showed a right-skewed pattern. Based on this distribution, we categorized hospitals into 4 groups based on annual patient volume. These were small (≤500), medium (501-1250), large (1251-2500), and major (2500). Within a hospital, task AUCs were averaged with equal weight, and hospitals were then weighted equally within their category. The paired difference was the AUC of the embedding minus that of one-hot encoding in the same hospital. To evaluate the association of data volume on performance, each hospital's paired difference was regressed on log_2 annual volume.

We also examined performance within patient strata defined by the number of documented *ICD-10* injury codes (1, 2-3, and 4 or more) and by ISS category (0-8, 9-15, 16-24, and 25 or more). Differences were then averaged and bootstrapped over hospitals for the hospital-specific models and over individual 2019 patients for the full NTDB models.

Prediction Model Hyperparameter Optimization

To permit prediction model hyperparameters to vary across data-availability scenarios (small, medium, large, major, and full NTDB), representations (one-hot encoding or embeddings at each of the 6 dimensions), and estimators (logistic regression or LightGBM), we tuned each combination separately. We thus ran 70 separate hyperparameter searches, each with 100 trials, using a tree-structured Parzen estimator sampler [21]. The 2018 development data were partitioned into 67% fitting and 33% validation sets, held fixed across searches. For hospital-specific models, each trial fitted one model per task and evaluable hospital. AUC was averaged first across tasks within a hospital, then across hospitals, weighting each hospital equally. For full NTDB models, AUC was averaged equally across tasks. The optimization objective was maximizing the mean AUC. Search spaces are reported in Tables S3 and S4 in Multimedia Appendix 1.

Embedding Space Visualization

To visualize the structure of the learned space, the 64-dimensional code embeddings were projected into 2 dimensions by uniform manifold approximation and projection with cosine distance, 30 neighbors, and a minimum distance of 0.1 [24]. The projection served visualization only and was not used in analysis.

Ethical Considerations

This study used data from the NTDB, which is released in deidentified form under a data use license following institutional application and fee. Data management was conducted in accordance with the NTDB data use agreement, with no ability to identify individual patients or centers.

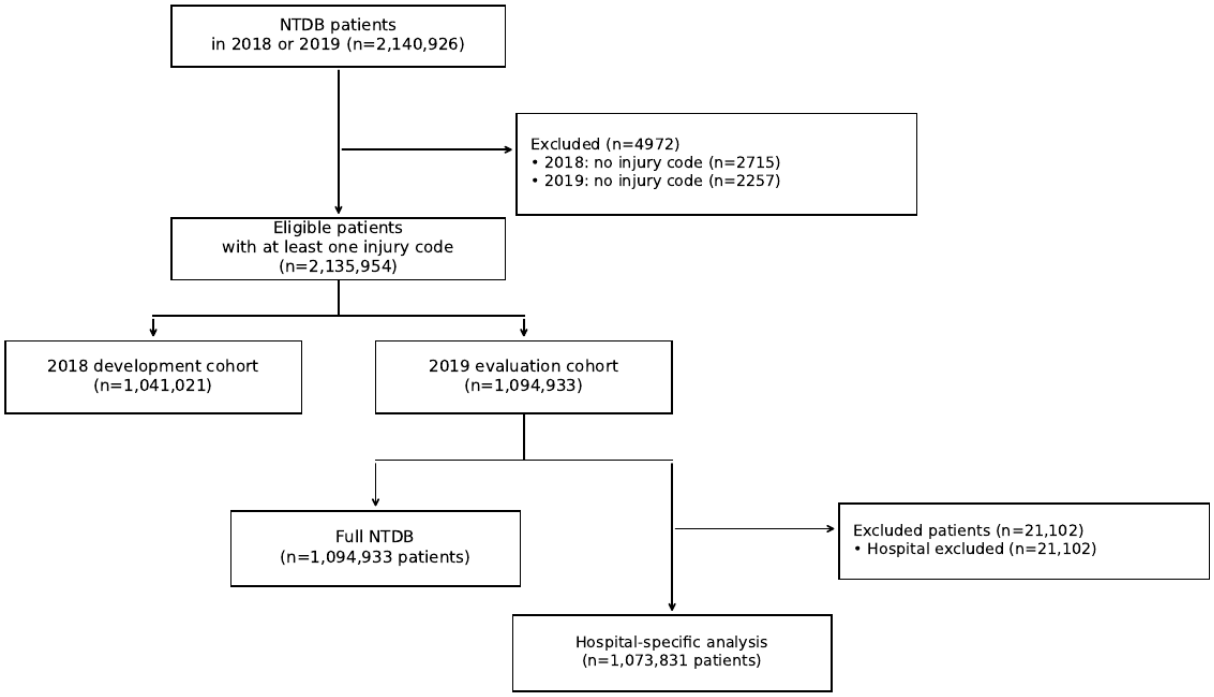
Results

Dataset Characteristics

The study included 2,140,926 patients with traumatic injuries from the NTDB in 2018 and 2019 (Figure 1). We excluded 4972 patients without a traumatic-injury *ICD-10* code, leaving 1,041,021 patients in the 2018 development cohort and 1,094,933 in the 2019 evaluation cohort. The full NTDB analyses used the complete development and evaluation cohorts without further exclusions. Of the 804 hospitals reporting in at least 1 year, 40 reported only in 2018, and 38 only in 2019. Of the 726 hospitals reporting in both years, 724 were included in the hospital-specific analyses. Two hospitals were excluded due to lack of prediction task data (Figure S2 in Multimedia Appendix 1). The included hospitals contributed 1,022,768 eligible patients for model development in 2018 and 1,073,831 for evaluation in 2019.

XSL•FO
RenderX

Figure 1. Patient inclusion flow for model development and evaluation. Patients with at least one traumatic-injury International Classification of Diseases, Tenth Edition code were included. The 2018 cohort was used for embedding model development and downstream-model training. The 2019 cohort was used for full National Trauma Data Bank (NTDB) and hospital-specific evaluation. Hospital inclusion for the hospital-specific analysis is detailed in Figure S2 in Multimedia Appendix 1.



Demographic and clinical characteristics were similar between the development and evaluation cohorts (Table 1). Most patients were male (1,261,503/2,135,675, 59%), and the largest age group comprised patients aged 25-49 years at 27% (548,978/1,997,155). Injuries were predominantly mild to moderate; of all patients, 50% (1,073,816/2,131,245) had an ISS below 9 and 34% (729,714/2,131,245) an ISS of 9-15, and 92% (1,860,550/2,014,911) presented with a GCS score of 13-15. A single injury code was documented for 28% (608,057/2,135,954) of patients, 2 or 3 for 38% (805,391/2,135,954), and 4 or more for 34% (722,506/2,135,954). End point–specific distributions, missing counts, and class distributions are reported in Table S5 in Multimedia Appendix 1.

The hospitals included were categorized into 4 size groups based on their yearly patient volume (Table S6 in Multimedia Appendix 1). The groups comprised small (n=120, 17%), medium (n=268, 37%), large (n=237, 33%), and major (n=99, 14%). Injury severity rose with hospital volume (Table S6 in Multimedia Appendix 1). The proportion of patients with an ISS of 16 or above increased from 9% (3805/40,542) at small hospitals to 18% (61,872/337,420) at major hospitals, and the proportion with a GCS score of 3-8 increased from 3% (1240/37,574) to 7% (21,813/322,042). The number of documented injury codes also rose. A single code was recorded for 41% (16,495/40,696) of patients at small hospitals, 33% (76,277/228,080) at medium, 27% (113,723/416,130) at large, and 24% (82,619/337,862) at major hospitals, and 4 or more codes for 21% (8654/40,696), 27% (62,212/228,080), 35% (145,711/416,130), and 39% (132,075/337,862) of patients in the same categories. From small to major hospitals, the proportion of women fell from 47% (18,964/40,690) to 38% (127,555/337,824) and the proportion aged 65 years or older from 39% (14,529/36,995) to 28% (91,277/321,590). No small hospitals were evaluable for hemorrhage-control surgery, due to lack of outcome data, so the small-hospital counts covered 15 rather than 16 tasks.

The number of unique *ICD-10* injury codes documented at a single hospital in 2018, which sets the input dimensionality of a hospital-specific one-hot encoding model, rose with hospital volume. The median was 200 (IQR 177-240) codes at small hospitals, 305 (IQR 259-337) at medium, 399 (IQR 371-429) at large, and 475 (IQR 449-503) at major hospitals.

The descriptive 2-dimensional projection of the 803 code embeddings is shown in Figure 2.

XSL•FO
RenderX

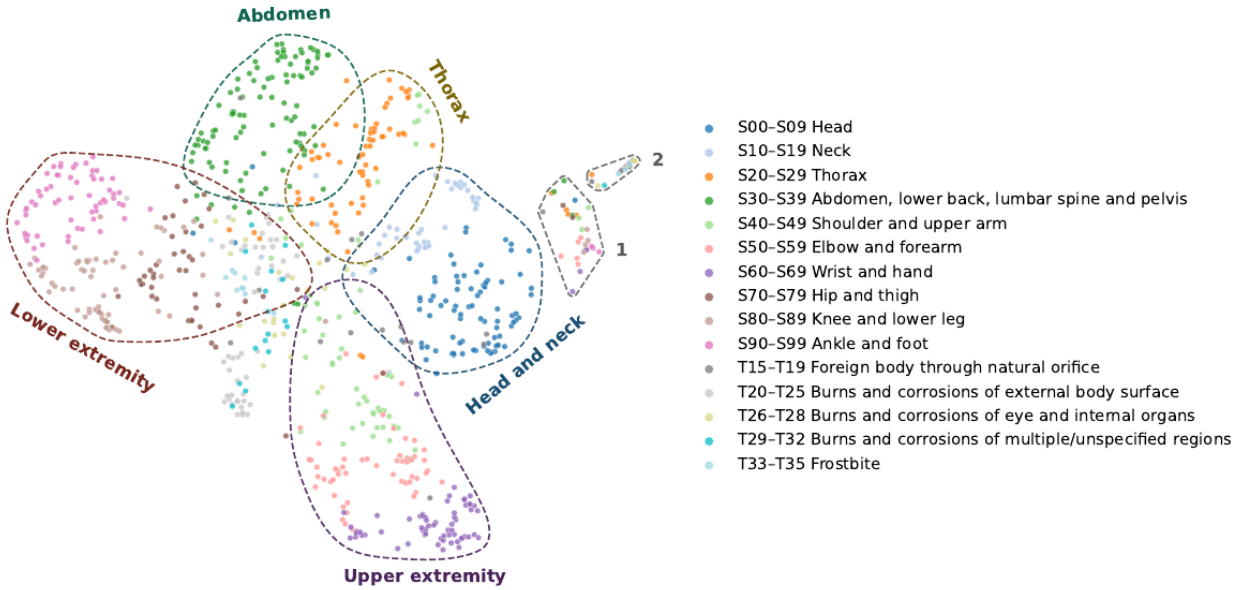
Table 1. Characteristics of the 2018 development and 2019 evaluation cohorts. Percentages use nonmissing observations.

Characteristic	2018	2019	Combined
Patients, n	1,041,021	1,094,933	2,135,954
Age (years), n (%)			
<18	112,239 (12)	118,181 (12)	230,420 (12)
18-24	89,039 (9)	88,706 (9)	177,745 (9)
25-49	270,929 (28)	278,049 (27)	548,978 (27)
50-64	185,177 (19)	192,487 (19)	377,664 (19)
65-79	188,522 (19)	207,721 (20)	396,243 (20)
≥80	128,258 (13)	137,847 (13)	266,105 (13)
Missing	66,857 (6)	71,942 (7)	138,799 (6)
Sex, n (%)			
Female	424,223 (41)	449,949 (41)	874,172 (41)
Male	616,677 (59)	644,826 (59)	1,261,503 (59)
Missing	121 (<1)	158 (<1)	279 (<1)
Glasgow Coma Scale, n (%)			
3-8	55,318 (6)	54,910 (5)	110,228 (5)
9-12	21,742 (2)	22,391 (2)	44,133 (2)
13-15	907,066 (92)	953,484 (93)	1,860,550 (92)
Missing	56,895 (5)	64,148 (6)	121,043 (6)
Injury Severity Score, n (%)			
0-8	522,977 (50)	550,839 (50)	1,073,816 (50)
9-15	353,657 (34)	376,057 (34)	729,714 (34)
16-24	96,697 (9)	99,184 (9)	195,881 (9)
≥25	65,037 (6)	66,797 (6)	131,834 (6)
Missing	2653 (<1)	2056 (<1)	4709 (<1)
Number of *ICD-10*[a] injury codes, n (%)			
1	295,283 (28)	312,774 (29)	608,057 (28)
2-3	392,168 (38)	413,223 (38)	805,391 (38)
≥4	353,570 (34)	368,936 (34)	722,506 (34)
Missing	0 (0)	0 (0)	0 (0)

[a]*ICD-10*: *International Classification of Diseases, Tenth Revision.*

XSL•FO
RenderX

Figure 2. Uniform manifold approximation and projection of the 64-dimensional embeddings for the 803 International Classification of Diseases, Tenth Edition (ICD-10) injury codes observed in the 2018 data. Colors indicate ICD-10 injury blocks. Colored dashed outlines mark where the bulk of the codes of each anatomical region lie. The anatomical regions are head and neck (S00-S19), thorax (S20-S29), abdomen (S30-S39), upper extremity (S40-S69), and lower extremity (S70-S99). Two additional regions are outlined in grey. Region 1 contains predominantly superficial injuries and contusions across anatomical sites, including contusions of the thorax (S20.2), shoulder and upper arm (S40.0), and elbow (S50.0). Region 2 contains rare codes, each recorded in at most 5 patients used to fit the embedding model.



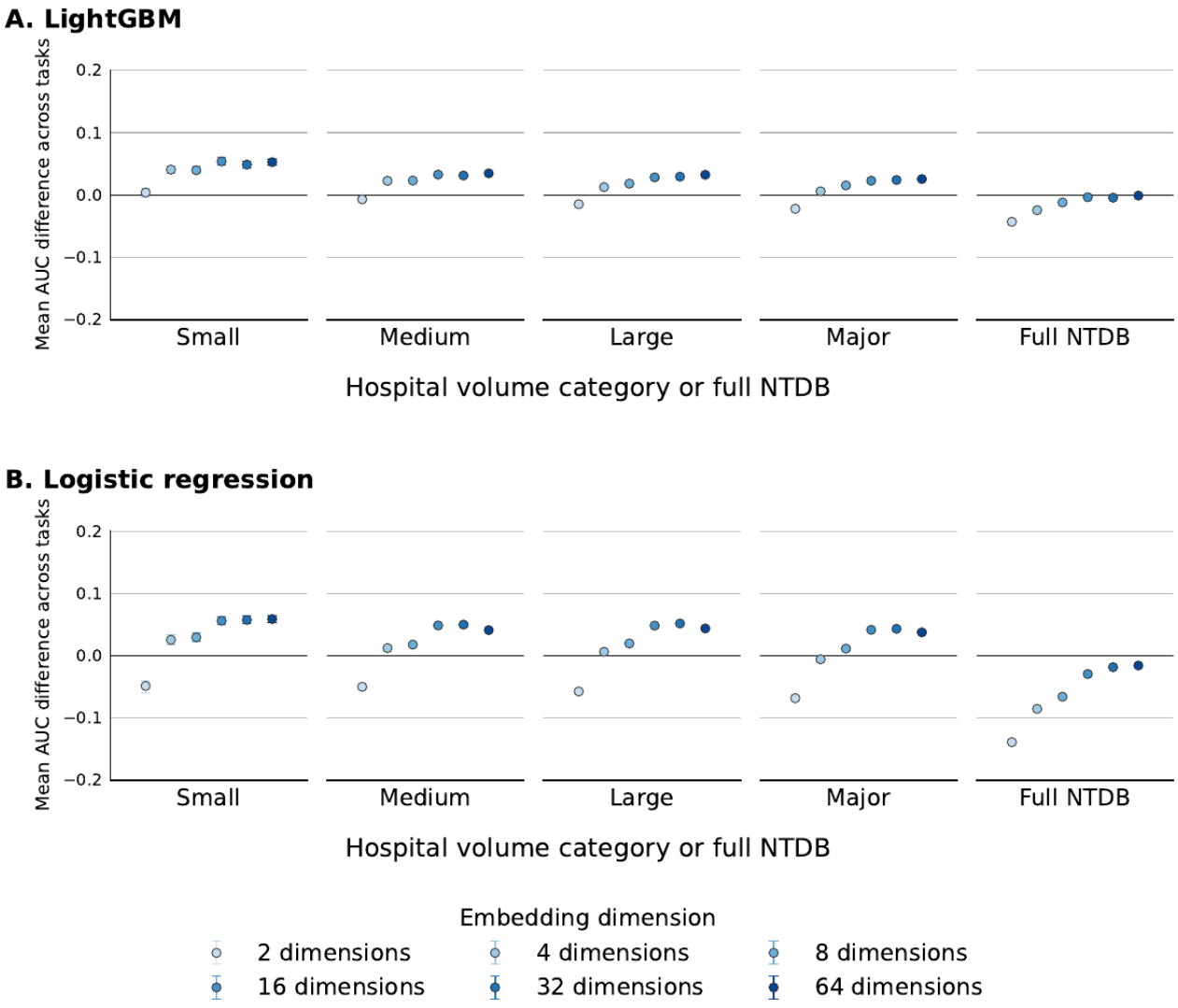
Performance of ICD-10 Injury Embeddings in LightGBM

For the LightGBM models, when aggregating results across evaluable prediction tasks, embeddings of 4, 8, 16, and 32 dimensions all had higher mean AUC than one-hot encoding in every hospital-volume category (Figure 3; Table S7 in Multimedia Appendix 1). For example, 16-dimensional embeddings exceeded one-hot encoding by 0.054 AUC (95% CI 0.048-0.060) in small hospitals, 0.033 (95% CI 0.030-0.035) in medium hospitals, 0.028 (95% CI 0.026-0.030) in large hospitals, and 0.023 (95% CI 0.020-0.025) in major hospitals. Two-dimensional embeddings performed worse than one-hot encoding in medium, large, and major hospitals. Treating annual volume as a continuous variable, the paired difference for embeddings of size 16 decreased by 0.009 (95% CI 0.008-0.010; $P<.001$) AUC per doubling of volume. The corresponding results for every dimension are reported in Table S8 in Multimedia Appendix 1.

In the full NTDB, the mean paired AUC difference at 16 dimensions was –0.004 (95% CI –0.004 to –0.003). Task-specific differences ranged from –0.016 (95% CI –0.018 to –0.014) for hemorrhage-control surgery to 0.003 (95% CI 0.002-0.005) for injury intent (Figure 4; Table S11 in Multimedia Appendix 1). Classifying each evaluable task by whether the 95% CI of its paired AUC difference excluded zero, 11 to 15 tasks favored 16-dimensional embeddings in every hospital-volume category, and at most one favored one-hot encoding. In the full NTDB at 16 dimensions, 5 tasks favored embeddings, 2 were inconclusive, and 9 favored one-hot encoding. At 64 dimensions, 7 tasks favored embeddings, 2 were inconclusive, and 7 favored one-hot encoding, with a mean paired AUC difference of –0.001 (95% CI –0.001 to –0.001). The counts in these 3 categories for every dimension, downstream model, and evaluation group are given in Table S9 in Multimedia Appendix 1, and the task-specific AUCs and paired differences at every embedding dimension are shown in Figures S3 and S4 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 3. Mean area under the receiver operating characteristic curve (AUC) difference across tasks between embeddings of 2, 4, 8, 16, 32, and 64 dimensions and one-hot encoding by hospital-volume category and in the full National Trauma Data Bank (NTDB). (A) LightGBM. (B) Logistic regression. Positive values favor the embeddings. Error bars indicate 95% CIs.

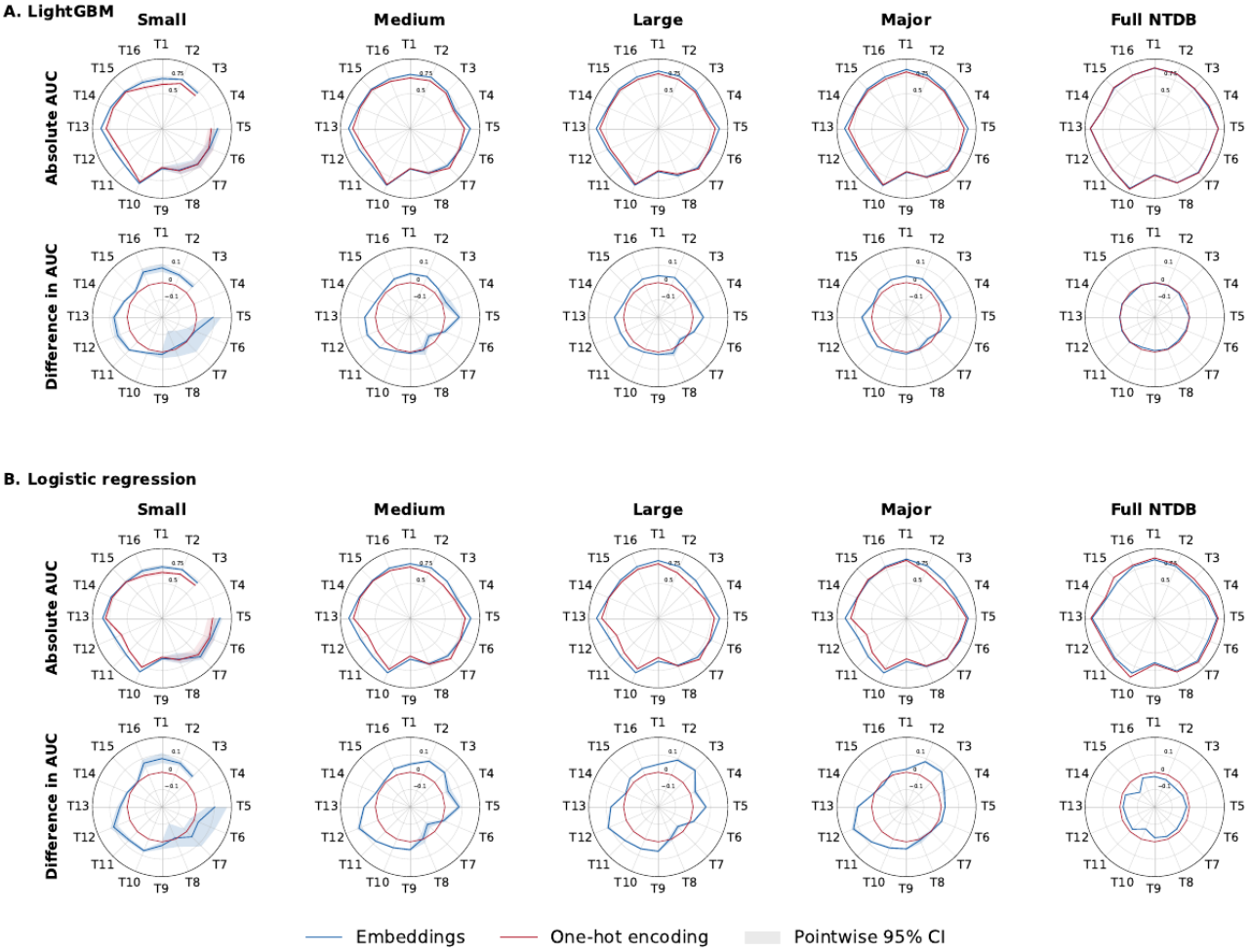


In the commonly modeled outcomes, in-hospital mortality, emergency department disposition, and blood transfusion within 4 hours, the hospital-specific 16-dimensional embeddings improved discrimination in each (Figure 4; Table S11 in Multimedia Appendix 1). For in-hospital mortality, AUC was 0.718 with the embeddings, compared with 0.634 with one-hot encoding in small hospitals, a difference of 0.084 (95% CI 0.061-0.107), and 0.849 against 0.813 in major hospitals, a difference of 0.036 (95% CI 0.032-0.041). For emergency department disposition, the corresponding values were 0.763 against 0.701 (difference 0.062, 95% CI 0.051-0.074) and 0.821 against 0.782 (0.039, 95% CI 0.035-0.043). Blood transfusion within 4 hours yielded an AUC of 0.801 against 0.708 in small hospitals (0.093, 95% CI 0.049-0.141), 0.865 against 0.783 in medium hospitals (0.082, 95% CI 0.072-0.095), and 0.886 against 0.831 in major hospitals (0.055, 95% CI 0.050-0.059). In the full NTDB, embeddings had lower AUC for mortality and emergency department disposition and higher AUC for blood transfusion. The corresponding AUCs were 0.870 against 0.871 for mortality (difference –0.001, 95% CI –0.002 to 0.000), 0.847 against 0.848 for emergency department disposition (–0.002, 95% CI –0.002 to –0.001), and 0.916 against 0.915 for blood transfusion (difference 0.001, 95% CI 0.000-0.002).

Stratifying patients by recorded injury-code count and by ISS category, the hospital-specific 16-dimensional embeddings had higher mean AUC than one-hot encoding in every stratum. In major hospitals the difference was 0.052 (95% CI 0.046-0.058) for patients carrying one code and 0.015 (95% CI 0.013-0.018) for those carrying 4 or more, and 0.042 (95% CI 0.037-0.049) for an ISS of 0-8 and 0.010 (95% CI 0.007-0.015) for an ISS of at least 25. In the full NTDB models the difference was 0.009 (95% CI 0.008-0.010) for patients carrying one code and –0.012 (95% CI –0.013 to –0.011) for those carrying 4 or more, and 0.003 (95% CI 0.001-0.005) for an ISS of 0-8 and –0.014 (95% CI –0.015 to –0.014) for an ISS of at least 25. At 64 dimensions, these differences narrowed to –0.007 (95% CI –0.008 to –0.007) for patients carrying 4 or more codes and –0.010 (95% CI –0.011 to –0.009) for an ISS of at least 25. Differences in every stratum, dimension, and evaluation group are given in Table S10 in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 4. Task-specific area under the receiver operating characteristic curve (AUC) for 16-dimensional embeddings and one-hot encoding, by prediction model and evaluation group. (A) LightGBM. (B) logistic regression. Within each panel, the upper row shows AUC and the lower row shows the paired AUC difference (embeddings minus one-hot encoding). Shaded regions indicate pointwise 95% CIs. NTDB: National Trauma Data Bank; T1: mortality; T2: emergency department disposition; T3: hospital disposition; T4: hemorrhage-control surgery; T5: 4-hour transfusion; T6: withdrawal of life-sustaining treatment; T7: cerebral monitoring; T8: Glasgow Coma Scale ≤8; T9: intensive care unit length of stay; T10: Injury Severity Score category; T11: mechanism; T12: intent; T13: trauma type; T14: computed tomography; T15: any operative procedure; T16: endotracheal airway insertion.



Performance of ICD-10 Injury Embeddings in Logistic Regression

For the logistic regression models, when aggregating results across evaluable prediction tasks, embeddings of 8 dimensions and above had higher mean AUC than one-hot encoding in every hospital-volume category (Figure 3; Table S7 in Multimedia Appendix 1). The 16-dimensional embeddings exceeded one-hot encoding by 0.056 AUC (95% CI 0.050-0.063) in small hospitals, 0.048 (95% CI 0.046-0.051) in medium hospitals, 0.048 (95% CI 0.046-0.051) in large hospitals, and 0.041 (95% CI 0.039-0.043) in major hospitals. Two-dimensional embeddings had lower mean AUC than one-hot encoding across all hospital-volume categories. Four-dimensional embeddings had higher mean AUC in small, medium, and large hospitals and lower mean AUC in major hospitals. Treating annual volume as a continuous variable, the paired difference at 16 dimensions was –0.005 AUC per doubling of volume (95% CI –0.006 to –0.003, P<.001). The corresponding results for every dimension are reported in Table S8 in Multimedia Appendix 1.

In the full NTDB, the mean paired AUC differences at 16 and 64 dimensions were –0.030 (95% CI –0.030 to –0.029) and –0.016 (95% CI –0.016 to –0.015), respectively. Classifying each evaluable task by whether the 95% CI of its paired AUC difference excluded zero, 11 to 14 tasks favored 16-dimensional embeddings in every hospital-volume category, and at most one favored one-hot encoding. In the full NTDB, all 16 tasks favored one-hot encoding at that dimension. The disadvantage persisted as the embeddings widened, with 0 tasks favoring embeddings, 1 inconclusive task, and 15 tasks favoring one-hot encoding at 64 dimensions (Table S9 in Multimedia Appendix 1). The task-specific AUCs and paired differences at every embedding dimension are shown in Figures S5 and S6 in Multimedia Appendix 1.

For the commonly modeled outcomes, in-hospital mortality, emergency department disposition, and blood transfusion within 4 hours, the hospital-specific 16-dimensional embeddings improved discrimination in each (Figure 4; Table S11 in Multimedia Appendix 1). For in-hospital mortality, AUC was 0.737 with the embeddings, compared with 0.658 with one-hot encoding in small hospitals, a difference of 0.078 (95% CI 0.053-0.110), and 0.850 against 0.833 in major hospitals, a

XSL•FO
RenderX

difference of 0.017 (95% CI 0.013-0.021). Blood transfusion within 4 hours had an AUC of 0.832 against 0.728 in small hospitals (0.104, 95% CI 0.061-0.175), 0.873 against 0.792 in medium hospitals (0.081, 95% CI 0.070-0.095), and 0.887 against 0.865 in major hospitals (0.022, 95% CI 0.019-0.026). Emergency department disposition had a difference of 0.072 (95% CI 0.057-0.086) in small hospitals, 0.092 (95% CI 0.086-0.098) in large hospitals, and 0.082 (95% CI 0.076-0.091) in major hospitals. In the full NTDB, the models using embeddings had lower AUC for all 3 outcomes. The corresponding AUCs were 0.841 against 0.866 for mortality (difference –0.025, 95% CI –0.026 to –0.024), 0.805 against 0.836 for emergency department disposition (–0.031, 95% CI –0.031 to –0.030), and 0.896 against 0.912 for blood transfusion (–0.016, 95% CI –0.017 to –0.015).

Stratifying patients by recorded injury-code count and by ISS category, the hospital-specific 16-dimensional embeddings had higher mean AUC than one-hot encoding in every stratum. In major hospitals the difference was 0.025 (95% CI 0.019-0.030) for patients carrying one code and 0.030 (95% CI 0.028-0.033) for those carrying 4 or more, and 0.037 (95% CI 0.030-0.042) for an ISS of 0-8 and 0.018 (95% CI 0.014-0.022) for an ISS of at least 25. In the full NTDB models the difference was –0.046 (95% CI –0.048 to –0.044) for patients carrying one code and –0.033 (95% CI –0.033 to –0.032) for those carrying 4 or more, and –0.043 (95% CI –0.045 to –0.040) for an ISS of 0-8 and –0.040 (95% CI –0.041 to –0.039) for an ISS of at least 25 (Table S10 in Multimedia Appendix 1).

Discussion

Principal Findings

We developed an *ICD-10* embedding model for traumatic injuries and found that low-dimensional representations retained information sufficient to support discrimination across multiple prediction tasks in substantially fewer dimensions than one-hot encoding. At the hospital level, embeddings of at least 4 dimensions with LightGBM and 8 dimensions with logistic regression achieved higher mean AUC than one-hot encoding in every hospital-volume category. For both model families, the largest gains were observed in the lowest-volume hospitals. In the full NTDB, one-hot encoding had higher mean AUC at every dimension with both models. With LightGBM, the mean loss in AUC narrowed from 0.012 at 8 dimensions to 0.001 at 64 dimensions, at which point embeddings had higher AUC on 7 tasks, 2 were inconclusive, and lower AUC on 7. Although these tallies do not convey effect magnitude, they show that the mean AUC reduction of 0.001 did not reflect uniformly poorer performance: embeddings improved discrimination on some tasks. With LightGBM, embedding performance relative to one-hot encoding declined as code count and injury severity increased. Widening the embeddings to 32 and 64 dimensions lessened this pattern, although the overall trend persisted. With logistic regression, performance relative to one-hot encoding remained generally stable as code count increased, but tended to decline among the patients with most severe injuries.

One possible explanation for the hospital-specific results is that the embeddings encoded structure learned from injury-code co-occurrence in the larger development dataset. In a one-hot representation, each injury code is represented by a separate indicator, and the downstream model must estimate its association with the outcome from patients in whom that code occurs. A code not observed during hospital-specific downstream-model training therefore has no fitted one-hot association and contributes no information to predictions.

In contrast, embeddings allow information to be shared across codes through relationships learned from their co-occurrence in the training data. A hospital-specific model can therefore represent a code absent from its own training data, given that the code was observed during embedding training. This mechanism is likely to be important in low-volume trauma centers, where direct modeling of individual one-hot codes becomes unreliable because of data sparsity. The hospital-level gains may therefore reflect 3 related advantages: reduced dimensionality, parameter sharing across codes learned from co-occurrence data, and representation of codes absent from the hospital-specific training data but observed in the evaluation data.

In the full NTDB evaluation, the benefits of unseen-code representation and information sharing across codes did not apply. This was most likely because the one-hot prediction models were trained on all eligible 2018 patients, as were the embedding models, and included a feature for every injury code observed that year. The embeddings therefore did not introduce injury information not already available to the one-hot models, but instead solely offered a compact representation of the same injury information. The lack of these advantages may explain why the performance gains observed at the hospital level were not retained in the full NTDB evaluation. These findings suggest that, beyond dimensionality reduction, the value of the current injury embeddings lies in transferring information learned from a larger population to downstream models trained on more limited data.

The structure learned by these embeddings arises from patterns of injury co-occurrence. No information on injury anatomy was provided to the model, yet codes from the same anatomical region lay close together in the embedding space (Figure 2), likely reflecting the tendency of injuries within the same anatomical region to co-occur. Mapping *ICD-10-CM* to WHO *ICD-10* removed laterality and reduced specificity regarding anatomical location and injury type, potentially reducing the absolute discrimination in downstream tasks. However, this mapping allows for use in registries that use or can be mapped to the international WHO *ICD-10* classification. Neither representation included age, physiology, mechanism of injury, or other clinical variables. The reported AUCs should therefore be interpreted as a comparison of alternative injury-code representations rather than as benchmarks against published clinical prediction models for the same outcomes.

Comparison With Prior Work

Our approach differs from previous methods for representing patients with embeddings in several respects. Most importantly, our embeddings are specific to patients with traumatic injuries. Compared with traditional injury summaries such as the ISS, it is data-driven and derives the patient representation directly

XSL•FO
RenderX

from individual injury codes. Unlike approaches that incorporate the *ICD-10* hierarchy into patient-derived embeddings or derive representations from code descriptions [25,26], our embeddings are learned solely from patterns of injury-code co-occurrence, without hierarchical or textual inputs. They therefore represent relationships between injury codes that arise from their observed co-occurrence in patients with traumatic injuries.

More general patient-embedding models have shown utility across clinical prediction tasks, but are not designed specifically to represent traumatic injuries and often require substantially broader clinical inputs [9,11,15,16,27]. Their representations are also considerably higher-dimensional, including 500 dimensions per patient in Deep Patient [9] and 768 in Apollo [15], compared with 8 to 64 dimensions in this study. By restricting the input to recorded injury codes, our approach produces a representation of traumatic injuries independently of medications, laboratory values, free text, longitudinal record history, or the temporal ordering of events. It can therefore be applied to trauma registries that record injuries using *ICD-10* codes without requiring additional clinical data.

Strengths and Limitations

This study has several strengths. First, the embedding model was developed in a large national trauma registry. The 2018 development data included 1,041,021 eligible patients from 766 hospitals and spanned a wide range of centers and case mixes. Second, the scale of the development data left few trauma codes unrepresented. The embedding model included 803 distinct injury codes, and of the 807 codes recorded in 2019, only 13 had not been observed during development, occurring in 32 of 1,094,933 evaluation patients. Third, development and evaluation were separated in time, with all model selection confined to 2018 and the 2019 data reserved for final evaluation. Fourth, the embeddings were evaluated in hospital-specific models across hospital-volume categories as well as in the full NTDB, allowing assessment both in smaller local datasets and at the scale of the national registry. Lastly, the representation was based on WHO *ICD-10* codes rather than the *ICD-10-CM* modification used in the United States, which may facilitate application in registries and health systems that do not use *ICD-10-CM*.

Several limitations should be considered. First, although the NTDB includes a diverse population of patients with traumatic injuries, it is not representative of trauma populations globally. It includes patients treated at hospitals in the United States, and differences in injury patterns and coding practices may limit generalizability to other health systems. Second, diagnosis and procedure codes may be finalized during or after admission, so codes can contain information that would not have been available at an earlier clinical decision point [28]. The evaluated tasks should therefore be interpreted as retrospective representation benchmarks rather than prospective clinical predictions. Furthermore, the procedure-code missingness indicator was retired after 2018, preventing identification of unknown procedure status in 2019. Missingness affected a small proportion of patients in 2018, suggesting that any resulting misclassification may have had a limited effect. Both representations used the same outcome labels; we therefore expect any resulting misclassification to have had a smaller effect on paired AUC differences than on absolute discrimination.

Third, injury codes not observed in the 2018 development data have no learned embedding and therefore cannot contribute directly to the representation. An unsupported code could instead be mapped to a supported ancestor in the *ICD-10* hierarchy. Fourth, the embedding dimensions are not directly interpretable. In contrast to one-hot encoding, in which each feature corresponds to a named injury code, the clinical meaning of an individual embedding dimension is not evident.

Fifth, the final embedding model was developed using the 2018 data, including patients who were subsequently used to fit the downstream models. However, no task-specific outcome labels were used during embedding training, which constitutes unsupervised pretraining rather than outcome leakage. Sixth, the relative performance of the embeddings was less favorable among the most severely injured patients. In hospital-specific models, the relative performance advantage of the embeddings generally diminished with increasing injury severity, although mean AUC remained higher than with one-hot encoding even among the most severely injured patients. In the full NTDB, LightGBM embeddings outperformed one-hot encoding among patients with mild injuries but underperformed among the most severely injured, while logistic-regression embeddings underperformed one-hot encoding in both severity strata. Transfers and other or unknown destinations were excluded from the corresponding disposition tasks. These exclusions may have disproportionately removed severely injured patients from lower-volume centers, limiting the generalizability of the observed absolute performance at the hospital level. Both representations were evaluated on the same patients; we therefore expect the effect of these exclusions on paired AUC differences to be limited.

Seventh, we identified no published embedding methods developed specifically for traumatic injuries. The embeddings were therefore compared only with one-hot encoding, without direct comparison with other embedding or dimensionality-reduction approaches. Lastly, downstream discrimination across 16 tasks was used as a proxy for preservation of clinically relevant injury information. Performance comparable to one-hot encoding suggests that the embeddings retained information relevant to these tasks, but does not exclude loss of information that was not required for any of them. Other potential uses of the representation, such as clustering or similarity search, were not evaluated.

Applications and Future Directions

Future research should validate the embeddings in trauma registries outside the United States, where injury patterns and coding practices may differ. Methods that better represent patients with many injury codes and those with the most severe injuries at lower embedding dimensions should also be investigated. In these groups, the advantage over one-hot encoding narrowed or reversed, although the difference was smaller at embedding dimensions of 32 or 64.

XSL•FO
RenderX

Potential applications include phenotyping and clustering of patients with traumatic injuries. These findings suggest that our injury-code embeddings may be considered during hyperparameter or model-development decisions in machine-learning pipelines. Even when embeddings only matched the discrimination of one-hot encoding, they did so using substantially fewer input features, with 8 to 16 values rather than 803 binary indicators. This reduces the dimensionality of the input presented to the downstream model and may reduce model complexity and computational requirements. Finally, with so few features, embeddings could serve as covariates when adjusting for injury pattern in statistical analyses, as embedding patient representations have previously been incorporated into methods for confounding adjustment in observational studies [29,30].

Conclusions

ICD-10 injury patterns can be represented in a low-dimensional embedding that preserves information relevant to downstream prediction. Compared with one-hot encoding, the embeddings achieved higher predictive performance in models trained at individual hospitals, particularly lower-volume hospitals. These gains may reflect reduced dimensionality, parameter sharing across codes, and representation of codes absent from hospital-specific training data, enabled by embeddings learned from the larger development cohort. In models trained on the full registry, one-hot encoding achieved higher mean AUC with both estimators, although the differences were minor (0.005 AUC difference) for LightGBM with 16- to 64-dimensional embeddings. These findings suggest that injury-code embeddings may provide a compact alternative to one-hot encoding, particularly when prediction models are fitted to more limited datasets. The representation requires only WHO *ICD-10* injury codes as input. The embeddings and accompanying software are openly available for further evaluation and application in trauma research.

Funding

No study-specific funding was received. The article processing charge is supported by the Swedish Bibsam Consortium.

Data Availability

The data analyzed in this study are not publicly redistributable. It was obtained from the American College of Surgeons National Trauma Data Bank Participant Use Files under a data-use agreement and is available from the American College of Surgeons subject to its application, fee, and licensing requirements [19]. Analysis code, trained models, and Python and R interfaces are available at GitHub [31].

Generative AI Use

Generative AI was used for language and code support. All generated outputs were reviewed. It was not used to generate the study's scientific ideas or results. The authors reviewed the manuscript and take responsibility for its content.

Authors' Contributions

KS and JB conceived and designed the study. All authors contributed to the methodology and development of the study protocol. JB performed data cleaning and pre-processing of the registry data. KS developed the machine learning pipeline and trained the embedding models. KS conducted the statistical analyses. KS and JB analyzed and interpreted the study results. KS wrote the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version. KS takes responsibility for the integrity and accuracy of the analysis in this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional data.
[PDF File (Adobe PDF File), 970 KB-Multimedia Appendix 1]

References

1. Gerdin M. The Risk of Dying: Predicting Trauma Mortality in Urban Indian Hospitals. Stockholm. Karolinska Institutet; 2015.
2. GBD 2023 Disease Injury Risk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet. 2025;406(10513):1873-1922. [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
3. Baker SP, O'Neill B, Haddon JW, Long WB. The Injury Severity Score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma. 1974;14(3):187-196. [doi: 10.1097/00005373-197403000-00001]

XSL•FO
RenderX

4. Zhang T, Nikouline A, Lightfoot D, Nolan B. Machine learning in the prediction of trauma outcomes: a systematic review. Ann Emerg Med. 2022;80(5):440-455. [doi: 10.1016/j.annemergmed.2022.05.011] [Medline: 35842343]
5. Salottolo K, Settell A, Uribe P, Akin S, Slone D, O'Neal E. The impact of the AIS 2005 revision on injury severity scores and clinical outcome measures. Injury. 2009;40(9):999-1003. [doi: 10.1016/j.injury.2009.0013]
6. Ghasemi P, Lee J. Unsupervised feature selection to identify important ICD-10 and ATC codes for machine learning on a cohort of patients with coronary heart disease: retrospective study. JMIR Med Inform. 2024;12:e52896. [FREE Full text] [doi: 10.2196/52896] [Medline: 39087585]
7. Osler T, Glance L, Buzas J, Mukamel D, Wagner JS, Dick A. A trauma mortality prediction model based on the Anatomic Injury Scale. Ann Surg. 2008;247(6):1041-1048. [doi: 10.1097/SLA.0b013e31816ffb3f] [Medline: 18520233]
8. Palmer CS, Cameron PA, Gabbe BJ. Comparison of revised functional capacity index scores with Abbreviated Injury Scale 2008 scores in predicting 12-month severe trauma outcomes. Inj Prev. 2020;26(2):138-146. [doi: 10.1136/injuryprev-2018-043085] [Medline: 30928915]
9. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094. [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
10. Shakya P, Khaneja A, Wagholikar KB. Predicting 30-days hospital readmission for patients with heart failure using electronic health record embeddings: comparative evaluation. JMIR Med Inform. 2025;13:e73020. [FREE Full text] [doi: 10.2196/73020] [Medline: 41288521]
11. Steiger E, Kroll LE. Patient embeddings from diagnosis codes for health care prediction tasks: pat2Vec machine learning framework. JMIR AI. 2023;2:e40755. [FREE Full text] [doi: 10.2196/40755] [Medline: 38875541]
12. Kim M, Kim Y, Kang HJ, Seo H, Choi H, Han J, et al. Leveraging BERT for embedding ICD codes from large scale cardiovascular EMR data to understand patient diagnostic patterns. BMC Med Inform Decis Mak. 2025;25(1):300. [FREE Full text] [doi: 10.1186/s12911-025-03145-x] [Medline: 40790571]
13. Zhou D, Tong H, Wang L, Liu S, Xiong X, Gan Z, et al. Representation learning to advance multi-institutional studies with electronic health record data from US and France. Nat Commun. 2026;17(1):4800. [FREE Full text] [doi: 10.1038/s41467-026-71152-1] [Medline: 41932876]
14. Gadd C, Gokhale K, Acharya A, Cooper J, Crowe F, Fitzsimmons L, et al. OPTIMAL Consortium. SurvivEHR: a competing risks, time-to-event foundation model for multiple long-term conditions from primary care electronic health records. NPJ Digit Med. 2026;9(1):546. [FREE Full text] [doi: 10.1038/s41746-026-02709-z] [Medline: 42106492]
15. Zhang A, Ding T, Wagner S, Tian C, Lu M, Pettit R. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale. arXiv. Preprint posted online on April 20, 2026. [doi: 10.48550/arXiv.2604.18570]
16. Wornow M, Thapa R, Steinberg E, Fries J, Shah N. EHRSHOT: an EHR benchmark for few-shot evaluation of foundation models. 2023. Presented at: Advances in Neural Information Processing Systems 36 (NeurIPS 2023); December 10-16; New Orleans, Louisiana. URL: https://tinyurl.com/bdfbzvwf [doi: 10.52202/075280-2933]
17. Guo LL, Fries J, Steinberg E, Fleming SL, Morse K, Aftandilian C, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. NPJ Digit Med. 2024;7(1):171. [FREE Full text] [doi: 10.1038/s41746-024-01166-w] [Medline: 38937550]
18. Ringdal KG, Coats TJ, Lefering R, Di Bartolomeo S, Steen PA, Røise O, et al. Utstein TCD expert panel. The Utstein template for uniform reporting of data following major trauma: a joint revision by SCANTEM, TARN, DGU-TR and RITG. Scand J Trauma Resusc Emerg Med. 2008;16:7. [FREE Full text] [doi: 10.1186/1757-7241-16-7] [Medline: 18957069]
19. National Trauma Data Bank. American College of Surgeons. 2026. URL: https://www.facs.org/quality-programs/trauma/quality/national-trauma-data-bank/ [accessed 2026-11-06]
20. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11:3371-3408. [FREE Full text] [doi: 10.1145/1390156.1390294]
21. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. 2019. Presented at: KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 4-8:2623-2631; Anchorage AK. [doi: 10.1145/3292500.3330701]
22. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. [doi: 10.3389/fninf.2014.00014]
23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. LightGBM: a highly efficient gradient boosting decision tree. 2017. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9; Long Beach, CA, USA.
24. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. Preprint posted online on February 9, 2018. [doi: 10.21105/joss.00861]
25. Finch A, Crowell A, Bhatia M, Parameshwarappa P, Chang Y, Martinez J, et al. Exploiting hierarchy in medical concept embedding. JAMIA Open. 2021;4(1):ooab022. [FREE Full text] [doi: 10.1093/jamiaopen/ooab022] [Medline: 33748691]
26. Kane MJ, King C, Esserman D, Latham NK, Greene EJ, Ganz DA. A compressed large language model embedding dataset of ICD 10 CM descriptions. BMC Bioinformatics. 2023;24(1):482. [FREE Full text] [doi: 10.1186/s12859-023-05597-2] [Medline: 38105180]

XSL•FO
RenderX

27. Beaney T, Jha S, Alaa A, Smith A, Clarke J, Woodcock T, et al. Comparing natural language processing representations of coded disease sequences for prediction in electronic health records. J Am Med Inform Assoc. 2024;31(7):1451-1462. [FREE Full text] [doi: 10.1093/jamia/ocae091] [Medline: 38719204]
28. Ramadan B, Liu M, Burkhart MC, Parker WF, Beaulieu-Jones BK. Diagnostic codes in AI prediction models and label leakage of same-admission clinical outcomes. JAMA Netw Open. 2025;8(12):e2550454. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.50454] [Medline: 41632159]
29. Weberpals J, Becker T, Davies J, Schmich F, Rüttinger D, Theis F, et al. Deep learning-based propensity scores for confounding control in comparative effectiveness research: a large-scale, real-world data study. Epidemiology. 2021;32(3):378-388. [doi: 10.1097/EDE.0000000000001338] [Medline: 33591049]
30. Gologorsky R, Somani SS, Neifert SN, Valliani AA, Link KE, Chen VJ, et al. Population scale latent space cohort matching for the improved use and exploration of observational trial data. Math Biosci Eng. 2022;19(7):6795-6813. [FREE Full text] [doi: 10.3934/mbe.2022320] [Medline: 35730283]
31. icd-10-injury-embeddings. GitHub. URL: https://github.com/noacs-io/icd-10-injury-embeddings [accessed 2026-11-13]

Abbreviations

AUC: area under the receiver operating characteristic curve
GCS: Glasgow Coma Scale
ICD: International Classification of Diseases
ICD-10: International Classification of Diseases, Tenth Edition
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
ISS: Injury Severity Score
NTDB: National Trauma Data Bank
WHO: World Health Organization

Edited by A Coristine; submitted 13.Jan.2026; peer-reviewed by GI Han, A Stey, S Liu; comments to author 10.Jun.2026; revised version received 02.Nov.2026; accepted 04.Nov.2026; published 23.Nov.2026

Please cite as:
Szolnoky K, Attergrim J, Ashfaq A, Linusson H, Gerdin Wärnberg M, Berg J
Development and Evaluation of Embeddings for Representing Injuries in Patients With Traumatic Injuries: Retrospective Cohort Study
JMIR Med Inform 2026;14:e91298
URL: https://medinform.jmir.org/2026/1/e91298
doi: 10.2196/91298
PMID:

©Kelvin Szolnoky, Jonatan Attergrim, Awais Ashfaq, Henrik Linusson, Martin Gerdin Wärnberg, Johanna Berg. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX