Clinical Relevance Score for Guided Trauma Injury Pattern Discovery with Weakly Supervised β-VAE

Qixuan Jin QIXUANJ@MIT.EDU
Massachusetts Institute of Technology, United States

Jacobien H.F. Oosterhoff J.H.F.OOSTERHOFF@TUDELFT.NL
Delft University of Technology, the Netherlands

Yepeng Huang YEPENGHUANG@HSPH.HARVARD.EDU
Harvard School of Public Health, United States

Marzyeh Ghassemi MGHASSEM@MIT.EDU
Massachusetts Institute of Technology, United States

Gabriel A. Brat GBRAT@BIDMC.HARVARD.EDU
Beth Israel Deaconess Medical Center & Harvard Medical School, United States

Abstract

Given the complexity of trauma presentations, particularly in those involving multiple areas of the body, overlooked injuries are common during the initial assessment by a clinician. We are motivated to develop an automated trauma pattern discovery framework for comprehensive identification of injury patterns which may eventually support diagnostic decision-making. We analyze 1,162,399 patients from the Trauma Quality Improvement Program with a disentangled variational autoencoder, weakly supervised by a latent-space classifier of auxiliary features. We also develop a novel scoring metric that serves as a proxy for clinical intuition in extracting clusters with clinically meaningful injury patterns. We validate the extracted clusters with clinical experts, and explore the patient characteristics of selected groupings. Our metric is able to perform model selection and effectively filter clusters for clinically-validated relevance.

Data and Code Availability We use data from the Trauma Quality Improvement Program (TQIP) for the years 2017 to 2019 (Committee on Trauma). This dataset of over 1 million patients in trauma centers across the US is available to researchers for "informational and research purposes" upon application through their website. The anonymized version of the code used in this paper can be accessed at `https://anonymous.4open.science/r/trauma_injury_clean-CF53`.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Trauma injury is one of the leading causes of death in the United States for the population under 45 years old (CDC, 2023). In 2020 alone, there were over 200,000 unintentional deaths, with unintentional falls and motor vehicle traffic accidents leading these statistics (CDC, 2021a). Trauma management is difficult because certain injuries may be more frequently overlooked despite standardized frameworks to assess trauma patients; approximately 15% to 22.3% of missed trauma injuries were clinically significant (Pfeifer and Pape, 2008). Primary and secondary surveys are carried out to assess and treat life-threatening injuries rapidly. Trauma programs often perform a tertiary survey to identify any missed injuries during the initial evaluation. Earlier identification of injuries can help avoid long-term injury and guide adequate treatment.

The medical literature has identified numerous trauma injuries that occur in groups or as *patterns*. For example, if a patient has a severe deceleration injury after a motor vehicle accident, along with an unstable "seatbelt" spine fracture, then the incidence of co-occurring intra-abdominal injuries can be as high as 89% (Tyroch et al., 2005). While there are injury patterns such as above that are well known in the clinical community, there has not been to date a comprehensive identification of trauma injury patterns. Historically, trauma pattern discovery has been an ad-hoc process based on the intuition and experience of a surgeon, primarily conducted within a small cohort at a single institution using classical statistics or rule-based methods. In addition to the low likelihood of identifying complex or rare patterns, the status quo suffers from small sample size and hospital and system bias, limiting the clinical relevance and generalizability of the identified patterns.

In this paper, we investigate the identification of injury patterns from the Trauma Quality Improvement Program (TQIP) – a large national trauma care database with over 1 million trauma patients collected from more than 875 participating trauma centers across the US (Committee on Trauma). Identification is a challenging problem as such patterns are unknown, and must be identified from retrospective sources for clinical validation. We pair an unsupervised disentangled variational autoencoder (β-VAE) with a multi-label classifier in the latent space, to efficiently create a latent space embedding. We use clinical diagnoses as input for reconstruction, and important clinical features such as age, mechanism of injury, correlation of the injury with mortality, and the Glasgow Coma Scale (GCS) as weak supervision. We use multilayer perceptron (MLP) classifiers to enforce self-organization in the latent space, i.e., such that groups with similar injury patterns will be clustered together.

After latent space clustering, we use a novel metric designed as a proxy for clinical relevance to extract injury patterns from identified subgroups. In clinical validation, our approach successfully identifies subgroups with known strongly associated patterns, such as the high occurrence of traumatic brain injury (TBI) in fall-related injuries with head injuries (Jager et al., 2000) and the combination of TBI and acetabular (hip-joint) fractures in motor vehicle collisions (Vella et al., 2017).

The main contributions of our work are:

1. We provide the first comprehensive exploration of injury patterns in the 2017-2019 TQIP patient cohort, developing a weakly supervised approach with auxiliary tasks that create a meaningful latent representation.

2. We demonstrate the utility of our proposed *clinical relevance* score (CR score) for effective model selection and filtering of clustering results.

3. We identify six subgroups of high clinical value, and analyze the associated injury patterns within the context of their typical mechanism of injury and other patient characteristics.

2. Related Works

2.1. Unsupervised and Weakly-supervised Representation Learning

Unsupervised representation learning seeks to identify patterns in the data without explicit labels. Weakly-supervised representation learning seeks to incorporate the added influence of some limited or imprecise signal. Specifically, we consider the β-VAE, a model proposed by Higgins et al. (2017) as a modification of the original VAE (Kingma and Welling, 2013) to encourage more disentanglement in the latent space. It has since been widely applied to automate the discovery of interpretable latent structures within data (Higgins et al., 2021; Li et al., 2021; Krajewski et al., 2018). The β-VAE is oftentimes enhanced with auxiliary features to learn in a weakly-supervised manner (Xie and Ma, 2019; Satheesh et al., 2021; Crespi et al., 2021; Sá and Roditi, 2021; Hsu and Lin, 2023). Our work seeks to identify clinically relevant trauma injury subgroups within the unlabeled data using the β-VAE framework augmented with weakly-supervised auxiliary classification features.

2.2. Domain-Guided Score

Evaluation of discovered clusters is difficult due to the lack of ground truth labels. Correctly interpreting the importance of a cluster oftentimes requires domain knowledge. Past works have addressed this challenge by incorporating scores measuring usefulness or relevance into clustering algorithms (Chang et al., 2017; Andreeva et al., 2020; Wenz et al., 2021). Specifically, Chang et al. (2017) used a polygenic risk score as the domain-specific score to guide their clustering of chronic obstructive pulmonary disease patients.

In the trauma care domain, the Glasgow Coma Scale (GCS), revised trauma score (RTS), injury severity score (ISS), and abbreviated injury scale (AIS) are widely used in the assessment of trauma patients during triage and the improvement of care (Lecky et al., 2014). However, most of the established scores require patient-specific assessment by clinical experts, which is time-consuming and laborious work for the experts. Many scores are also not capable of identifying co-occurring injuries. In light of the shortcomings of established scores, we develop our own cluster relevance score with the aid of domain experts (see Sec. 4.4). In contrast to prior works, we use the score as a means for evaluation and model selection, instead of a constraint during clustering.

2.3. Pattern Discovery in Traumatic Injuries

Trauma data is often high-dimensional and consists of complex and heterogeneous clinical and demographic information, making it difficult to identify meaningful patterns directly. The identification of trauma injury patterns is oftentimes conducted in cohorts with a small sample size ($\leq 5,000$), for a specific trauma patient population (DGU et al., 2015; Chichom-Mefire et al., 2018). For instance, association rule mining has also been used to identify 77 individual-based injury patterns in multi-trauma road users (Fagerlind et al., 2022). Ensemble classifiers have been used to detect vascular injury in trauma care (Metzger et al., 2015). The closest previous work to ours is the unsupervised mining of temporal injury patterns in the larger dataset of general trauma patients ($\sim$ 500,000) with restricted Boltzmann machine (Mehrabi et al., 2015). The focus of this work differs from ours, however, as we are not interested in the progression of the patient's condition over time. To the best of our knowledge, weakly-supervised representation learning has not been applied to conduct general injury pattern discovery in such a large, heterogeneous cohort of trauma patients.

3. Data and Preprocessing

3.1. Cohort Selection

We include patients from the 2017-2019 TQIP database. As we do not detect any significant temporal shift in the data, we aggregate the patient cohorts across the years (Section E.4). Patients are excluded when they (1) are younger than 16 years of age at the time of record, or (2) are reported dead on arrival. The selected cohort consists of $n =$ 1,162,399 patients, with a 78-22 split into train set ($n_{\text{train}} =$ 903,267) and test set ($n_{\text{test}} =$ 259,132). Each patient record is uniquely included in either set.

3.2. Preprocessing of Injury Codes and Auxiliary Features

An International Classification of Disease (ICD) code is a seven-character, globally used code to categorize disease. We truncate ICD-10 trauma codes (codes that start with an 'S' or a 'T') to the first four characters, thus including the highest level of their subcategory. Non-trauma ICD-10 codes are shortened to their main category (first three characters). After truncation, we have a set of 1,317 unique codes. The top 500 most frequent codes are selected, in order to exclude the more uncommon diagnoses (Section B). The highest frequency injury code is "Multiple fractures of ribs" with a prevalence of 15.9% and the lowest frequency code is "Hypertensive heart and chronic kidney disease" with a prevalence of 0.03%. For each patient, the set of diagnosed conditions is binarized to form the input feature vector x.

We refer readers to Appendix A for a detailed description of how the auxiliary features of age, the mechanism of injury, GCS, and high-risk injuries are preprocessed to the feature vector c. The mechanism of injury consists of eight categories describing the mechanism by which the patient was injured (Table 4). GCS is a measure of patient responsiveness and serves as a proxy for the degree of traumatic brain injury (Table 5). High-risk injuries are injuries that are highly associated with mortality, and are thus important to be identified early (Table 7).

4. Methods

Suppose we have a dataset of n patients, each with an input binary feature vector x of injury codes. We also have a normalized vector c of the auxiliary features per patient.

4.1. Disentangled Variational Autoencoder

We use a standard β-VAE framework (Higgins et al., 2017) for learning the latent representations of trauma injuries. The encoder compresses input information into a latent representation, which the decoder samples from and seeks to reconstruct the input. Specifically, distribution $q_\phi(z|x)$ encodes x to latent vector z, while distribution $p_\theta(x|z)$ decodes sampled z to reconstructed x_{recon}. ϕ and θ parameterize their respective distributions. We use an isotropic unit Gaussian $\mathcal{N}(0, 1)$ as the latent prior $p(z)$. VAEs are trained by maximizing the evidence lower bound

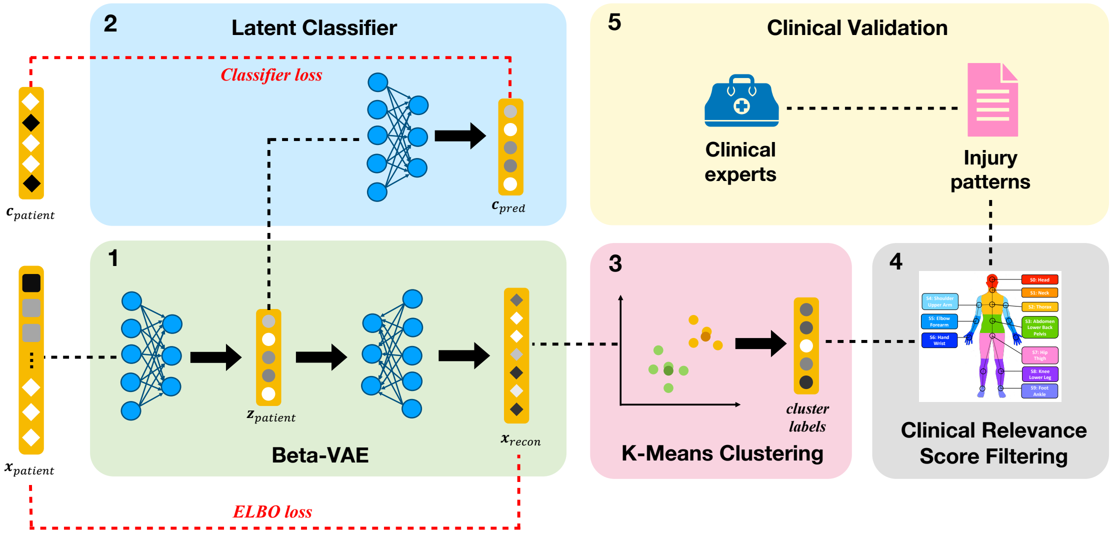


Figure 1: Our β-VAE Classifier framework for injury pattern discovery consists of five components: (1) the β-VAE learns a latent space, (2) a classifier of auxiliary features provides weak supervision, (3) clustering is performed on the latent space (4) the CR score selects clusters with clinically interesting injury patterns, (5) selected injury patterns are validated by clinical experts.

(ELBO). In practice, we train the model to minimize the objective function:

$$\text{ELBO} = -\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) + \beta D_{\text{KL}}\big(q_\phi(z|x)||p(z)\big)$$

We implement the encoder $q_\phi(z|x)$ as a two-layer MLP that learns the mean μ and variance σ of the distribution $q_\phi(z|x)$. We implement the decoder $p_\theta(x|z)$ as a two-layer MLP with sigmoid activation. Code uses Tensorflow 2.4.1 (Abadi et al., 2015). We set a latent representation of dimension 64. All models are trained for 100 epochs at a learning rate of 0.001.

4.2. Latent Space Classifier with Auxiliary Features

The auxiliary classifier $f_\psi(c|z)$, parameterized by ψ, predicts c from z. Classifier loss is implemented as binary cross-entropy and added to the ELBO loss with weight γ. The final loss function of the β-VAE Classifier model is:

$$\mathcal{L} = -\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) + \beta D_{\text{KL}}\big(q_\phi(z|x)||p(z)\big) \\ -\gamma\frac{1}{n}\sum_{i}^{n}\sum_{j}^{\dim(c)} c_{ij}\log\big(f_\psi(c_{ij}|z_{ij})\big)$$

4.3. Clustering

We use K-Means clustering with Euclidean distance and $K = 30$ as the number of clusters. We motivate this choice in Appendix E. We use the sklearn.clustering library for the implementation (Buitinck et al., 2013).

4.4. Clinical Relevance Score

The CR score consists of four submetrics.

Body-Spatial Submetric (bs) Trauma injury patterns that contain injuries that are spatially far apart and span multiple body regions are of greater clinical interest, since they are often indicative of complex injury patterns. Injuries are categorized into the ten anatomical regions given by the highest level of the ICD-10 hierarchy. We represent each region as a node in a connected graph. The body-spatial submetric $\mathbf{bs}(\cdot,\cdot)$ is the path length between two nodes, normalized to $[0, 1]$.

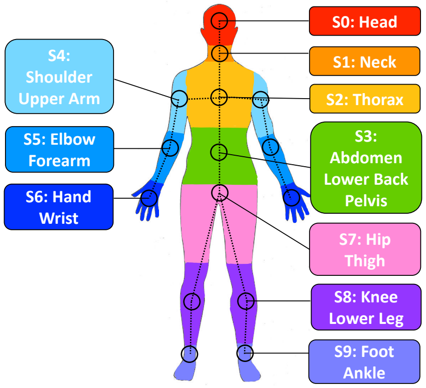


Figure 2: The anatomical graph for computing the body-spatial metric.

Internal-External Submetric (ie) External injuries can often serve as warning signs for important correlated internal injuries that can be detected early through a screening procedure. See Table A for the set of injuries assigned non-zero weights w_i. For a pair of injuries h_1 and h_2 with internal weights w_{i_1} and w_{i_2}, the metric is computed as:

$$\mathbf{ie}(h_1, h_2) = \begin{cases} \text{abs}(w_{i_1} - w_{i_2}) & \text{if } w_{i_1} \neq w_{i_2} \\ w_{i_1}/2 & \text{if } w_{i_1} = w_{i_2} \end{cases}$$

We also down-weight superficial injuries by returning $\mathbf{ie}(\cdot,\cdot) = -1$ if the pair contains a superficial injury.

High-Risk Submetric (hr) We refer to the same list of 50 high-risk injuries as defined in the auxiliary feature preprocessing (see Section A). Each high-risk injury (Table 7) adds 0.5 to the submetric $\mathbf{hs}(\cdot,\cdot)$.

Correlation (corr) We use the already computed Pearson correlation of the injury pairs for the output of the $\mathbf{corr}(\cdot,\cdot)$ function.

Algorithm for Computing CR Score Default hyperparameter values are noted in parentheses.

As input, we have cluster labels and patient injury codes. We include clusters with more than S_κ number of patients ($S_\kappa = 259$, or 0.1% of test cohort size). For a particular cluster, we select injuries that occur at a frequency higher than threshold κ ($\kappa = 0.04$). We compute the Pearson correlation between pairs of injuries. We select the top S_α pairs of injuries as ranked by correlation ($S_\alpha = 50$). We remove pairs with a correlation less than threshold α ($\alpha = 0.25$). For all pairs of injuries h_a and h_b, we compute the weighted score:

$$\begin{aligned} w = w_{bs} * \mathbf{bs}(h_a, h_b) + w_{ie} * \mathbf{ie}(h_a, h_b) \\ + w_{hr} * \mathbf{hr}(h_a, h_b) \end{aligned}$$

If the injury pair has $w > 0$, we add the correlation submetric and compute the final CR score as:

$$\text{CR score} = w + w_{corr} * \mathbf{corr}(h_a, h_b)$$

For all pairs with a positive CR score, we merge pairs with shared injuries into larger sets to form injury patterns. We average the CR score across all pairs to compute the CR score for the cluster, and average across all valid clusters to compute the CR score for the model.

We compute the CR score on the test cohort. **The default weights for the CR score submetric are: $w_{bs} = 0.5, w_{ie} = 0.2, w_{hr} = 0.2, w_{corr} = 0.1$.** These values are determined jointly with our collaborating clinicians, and reflect their preferences for what counts as “clinically meaningful”. We have the highest weight for the body-spatial submetric, since injury patterns that are further apart in the body are associated with more complex patterns. Internal-external relationships and high-risk injuries are given the same weight, while correlation is given the least weight.

4.5. Model Baselines

As baselines for the β-VAE Classifier framework, we consider the vanilla β-VAE without the classifier as well as singular value decomposition (SVD). SVD is a linear dimensionality reduction method that relies on matrix factorization. SVD is implemented with TruncatedSVD from the sklearn library (Buitinck et al., 2013). The chosen baselines are appropriate because SVD serves as a simple, but robust baseline, while the vanilla β-VAE models the completely unsupervised setting without auxiliary information.

4.6. Evaluation

4.6.1. Auxiliary Task Evaluation

To evaluate model performance in the auxiliary task, we use standard binary classification metrics such as the area under the ROC curve (AUC), the F1 score, recall, and precision (Hossin and Sulaiman, 2015).

4.6.2. Unsupervised Clustering Evaluation

For the evaluation of the unsupervised representations, we use two common metrics: the silhouette coefficient (Silh. Coef.) (Rousseeuw, 1987) and the Calinski-Harabasz Index (CH Index) (Calinski and Harabasz, 1974). The silhouette coefficient measures how well points group within their own cluster in comparison with neighboring clusters. The silhouette coefficient ranges from $[-1, 1]$, with a more positive score indicating denser and better-separated clusters. The CH index measures the ratio of between-cluster variance (larger better) and within-cluster variance (smaller better). A higher CH index indicates better clustering. Since we have the same dimension size and normalize the latent space before evaluation, the CH index is comparable across representations (Buitinck et al., 2013). We use these metrics as we may expect well-defined clusters with better unsupervised clustering metrics to also contain strongly associated injury patterns with higher clinical relevance.

Moreover, the clustering is also evaluated with our developed CR score and a manually generated Expert Rating with clinical experts.

5. Results

5.1. Model Performance on Auxiliary Classification Tasks

Metrics	SVD	BetaVAE	BetaVAE Classifier
AUC	0.793 (± 0.096)	0.821 (± 0.100)	**0.842** **(± 0.094)**
F1	0.401 (± 0.347)	0.432 (± 0.354)	**0.488** **(± 0.345)**
Recall	0.384 (± 0.363)	0.406 (± 0.366)	**0.457** **(± 0.367)**
Prec.	0.585 (± 0.269)	0.605 (± 0.284)	**0.618** **(± 0.283)**

Table 1: Model performance averaged on 12 auxiliary tasks across 5 randomized runs with 95% confidence intervals.

First, we evaluate the performance of the β-VAE Classifier model as compared to the baseline models on the supervised classification task of predicting auxiliary features from the latent space (Table 5.1).

Note, while auxiliary tasks are not the goal of our work, good performance indicates that weak supervision is helping the representations converge to meaningful spaces. We find that, on average, the β-VAE Classifier outperforms the β-VAE and SVD model (AUC 0.842 vs. 0.821 vs. 0.793) on all 12 tasks. See Appendix C for further details.

5.2. Evaluating Latent Space Clustering with CR Score

Next, we evaluate the learned latent space clusterings of the β-VAE Classifier model against baselines for the injury pattern discovery task. In terms of unsupervised clustering metrics, we see in Table 5.2 that the β-VAE Classifier performs best for the CH index, while the SVD model performs best for the silhouette coefficient. The higher silhouette coefficient for the SVD model may be explained by the presence of small, compact clusters at a considerable distance from the main cluster density (Figure 5). The clusters of the β-VAE Classifier are better separated than the clusters of the β-VAE, due to the auxiliary weak supervision. We note that the two clustering metrics *disagree* on the model type with the best clustering. This disagreement further motivates the need for a more direct clinical metric to evaluate cluster quality.

Metrics	SVD	BetaVAE	BetaVAE Classifier
CR Score	0.104 (± .003)	0.116 (± 0.020)	**0.140** **(± 0.003)**
Silh. Coef.	**0.123** **(± 0.007)**	0.043 (± 0.007)	0.064 (± 0.004)
CH Index	253.7 (± 10.7)	181.8 (± 11.6)	**327.3** **(± 18.2)**

Table 2: Evaluation of latent representations across 5 randomized runs with 95% confidence intervals.

In terms of the CR score (Table 5.2), the β-VAE Classifier (CR = 0.140) performs better than the vanilla β-VAE (CR = 0.116), which performs better than the SVD (CR = 0.104). The addition of the auxiliary signal seems to induce the learning of more meaningful clusters.

Now, we qualitatively validate that the CR score is able to capture the concept of clinical relevance better than the clustering metrics by examining the cluster output manually. We observe that head injuries tend to dominate the cluster composition in the SVD and β-VAE model (percent containing head injuries = 52.6% and 57.9%, respectively). This dominating effect is undesirable and is much less present in the β-VAE Classifier (percent containing head injuries = 22.7%). Instead, the clusters often span multiple body regions. We observe injury patterns such as {Fracture of thumb, Traumatic amputation of thumb} and {Injury of radial artery at wrist and hand level, Fracture of lower end of radius} for the β-VAE Classifier that we do not see in the baseline models. To summarize, VAE models capture more variety of injury patterns, and the β-VAE Classifier is able to capture patterns spanning more body regions than the vanilla β-VAE.

5.3. Validating the Model Selection Capacity of the CR Score

To illustrate the capacity of the CR score for nuanced model selection in addition to evaluation, we first train a pool of 50 candidate β-VAE Classifiers ($\beta = 5, \gamma = 1$). We then compare the clinical relevance of the model with the best silhouette coefficient and CH index ("Unsup Top Model") to the model with the best CR score ("CR Top Model"). We analyze the performance of these two models as "CR Top Model" is the best model according to our developed metric, while "Unsup Top Model" is the model we would have picked if we don't have access to the CR score.

Metrics	**Unsup Top Model**	**CR Top Model**
Silh. Coef.	0.092	0.037
CH Index	497.5	233.0
CR Score	0.144	0.168
Expert Rating	1.034	1.227

Table 3: Evaluation of latent representations of the best model by clustering metrics (Unsup Top Model) and the best model by the CR score (CR Top Model).

As shown in Table 5.3, the difference in the CR Score between the two models is discernible but not large. To assess whether this difference is clinically perceptible, a collaborating clinician (blinded to cluster source method) labeled the clusters on a scale from 0 to 2 based on how "clinically relevant and interesting" they believed each cluster to be. Higher is more clinically relevant. We average these scores per model to form the Expert Rating.

We find that the best model chosen by clustering metrics versus that chosen by CR have an Expert Rating of 1.034 versus 1.227 respectively (Table 5.3). The trend in the Expert Rating concurs with the trend in the CR score. Qualitatively, the clinicians also remarked that the Unsup Top Model has a higher proportion of clusters with expected injury patterns that would not be of interest, such as: {"Traumatic pneumothorax (collapsed lung)", "Multiple fractures of ribs"}, {"Fracture of lower end of ulna (forearm bone)", "Fracture of lower end of radius (forearm bone)"}, and {"Fracture of nasal bones", "Open wound of nose" }. See Section F.1 for additional analysis.

6. Tuning Submetrics to Customize CR Score

The submetric weights of the CR score in the previous result sections are tuned to roughly approximate the clinical intuition of our collaborating clinicians. A better understanding of the typical injury patterns favored by each submetric can aid in the tuning process of adapting the CR score to different clinical tasks. In this section, we explore the influence of each submetric on the extracted injury patterns. In each submetric section, we set the submetric to the maximum value. For instance, to analyze the body-spatial metric, we would define the weights ($w_{bs} = 1, w_{ie} = 0, w_{hr} = 0, w_{corr} = 0$).

Body-Spatial Submetric (bs) The body-spatial submetric favors clusters with complex injury patterns spanning multiple body regions. The further apart the injuries are located on the body, the higher the value will be. We observe that the top extracted clusters can concurrently span the head, thoracic, abdominal, and extremity regions. We also observe clusters with injury patterns such as {"Fracture of acetabulum (hip bone)", "Fracture of radius (forearm bone)", "Fracture of calcaneus (heel bone)"}. Although this fracture pattern only spans extremities, the injuries themselves are spatially far as the lower arm is far from the foot according to Figure 2.

Internal-External Submetric (ie) This submetric awards pairs of injuries in which one is external and one is internal. The top clusters as ranked by this submetric will almost certainly discover some variation of the injury patterns: {"Traumatic pneumothorax (collapsed lung)", "Multiple fractures of ribs"} and {"Fracture of base of skull", "Traumatic subdural hemorrhage (brain bleed)"}. Note that although both patterns contain internal and external injuries, the patterns themselves are not as clinically interesting due to how common and expected they are.

The utility of the internal-external submetric, however, lies in awarding patterns such as {"Injury of colon", "Injury of small intestine", "Injury of other intra-abdominal organs", "Injury of iliac blood vessels (abdominal)", "Fracture of ilium (pelvic bone)"}. Here, the fracture is the visible external presentation of the harder-to-detect internal injuries of the colon and other intra-abdominal organs. Discovering such injury patterns can aid with diagnosis as the presence of external injuries can alert the clinician to the potential co-occurrence of internal injuries characterized by the pattern.

High-Risk Submetric (hr) The utility of the high-risk submetric is fairly self-evident, as it is important to identify clusters that contain injuries highly correlated with patient mortality. The early detection of high-risk injuries can improve trauma management. A few injury pattern types that we typically observe when we rank by the high-risk submetric are:

1. *Severe head injuries*: typically some combination of cerebral edema (brain swelling), hemorrhage, and fracture of some part of the skull.
2. *Spine fractures*: typically fractures of two or more contiguous vertebrae, such as the fracture of the first and second cervical (neck) vertebra.
3. *Thoracic injuries*: typically some combination of traumatic hemopneumothorax (bleed in a collapsed lung), flail chest (unstable chest wall), and rib fractures.

Correlation (corr) If we rank by correlation, then we will mostly discover clusters with injuries that have expected associations. The pattern may characterize the injuries that are spatially near or are caused by a single, clear mechanism of injury. We obtain fractures of neighboring bones (e.g. acetabulum (hip) and pubis (pelvis)) and internal organs (e.g. kidney, pancreas, liver, gallbladder, and bile duct). Interestingly, we also observe a higher occurrence of milder severity injury patterns. For instance, we observe patterns of sprain injuries (e.g. sprain of collateral ligament (outer knee) of knee and tear of meniscus (inside the knee)) and of non-trauma-related conditions (e.g. essential hypertension (high blood pressure) and respiratory failure).

7. Case Study

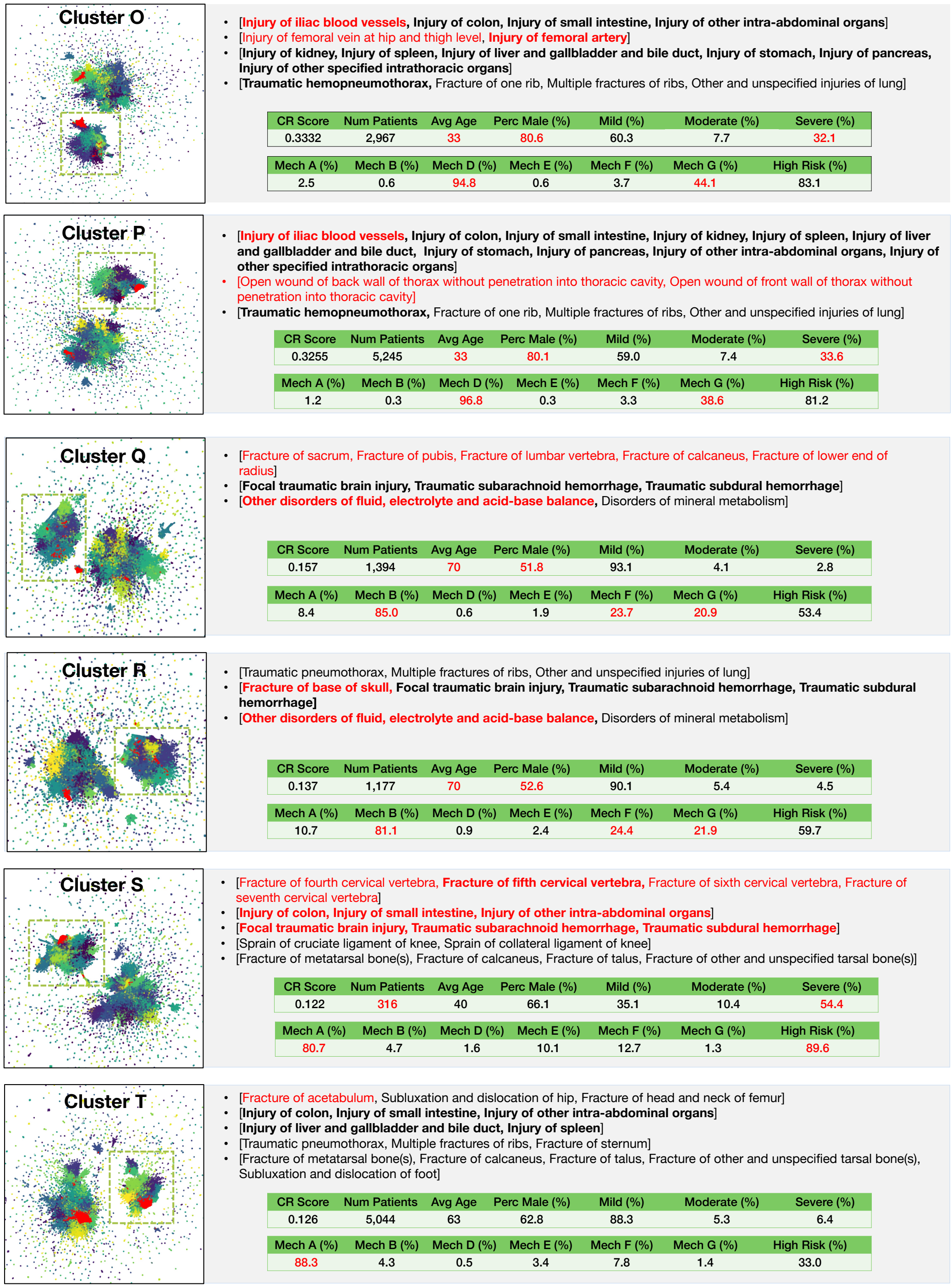
Now, with our tuned model and CR score, we provide a clinician-validated analysis of trauma injury subgroups in the TQIP 2017-2019 patient cohort. We present these clinical results as a demonstration of how our framework can be utilized to conduct retrospective analysis for the discovery of clinically relevant injury patterns.

7.1. Setup

For 50 randomized runs of the β-VAE Classifier, we aggregate all cluster outputs and ranked the clusters by the CR score. We present the ranked list to the clinician. The clinician identifies six subgroups of interest (see Figure 3). The six subgroups span three mechanisms of injury: penetrating trauma, falls, and motor vehicle accidents.

7.2. Penetrating Trauma Subgroups

Penetrating trauma (Mech D) is an open wound injury caused by a foreign object piercing the skin, such as gunshots or stab wounds injury (Fitch et al.). We identify **Cluster O** and **Cluster P** as two clusters that highlight the importance of the assessment of abdominal vascular trauma (blood vessel injury in the abdominal area). Abdominal vascular trauma is rare, but when it does occur, high mortality rates are seen up to 60% of all cases (Kobayashi et al., 2016). **Cluster O** includes an injury to the iliac and femoral vessels (abdominal blood vessels). **Cluster P** also represents an injury to the iliac vascularity. Iliac vessel injuries are uncommon, but among the most lethal and challenging injuries, and patients often arrive in shock secondary to massive blood loss (Kim et al., 2016). When abdominal vascular injury is suspected, immediate attempt to control the bleeding is essen-



Cluster O

- [**Injury of iliac blood vessels**, **Injury of colon, Injury of small intestine, Injury of other intra-abdominal organs**]
- [Injury of femoral vein at hip and thigh level, **Injury of femoral artery**]
- [**Injury of kidney, Injury of spleen, Injury of liver and gallbladder and bile duct, Injury of stomach, Injury of pancreas, Injury of other specified intrathoracic organs**]
- [**Traumatic hemopneumothorax,** Fracture of one rib, Multiple fractures of ribs, Other and unspecified injuries of lung]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.3332	2,967	33	80.6	60.3	7.7	32.1

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
2.5	0.6	94.8	0.6	3.7	44.1	83.1

Cluster P

- [**Injury of iliac blood vessels**, **Injury of colon, Injury of small intestine, Injury of kidney, Injury of spleen, Injury of liver and gallbladder and bile duct, Injury of stomach, Injury of pancreas, Injury of other intra-abdominal organs, Injury of other specified intrathoracic organs**]
- [Open wound of back wall of thorax without penetration into thoracic cavity, Open wound of front wall of thorax without penetration into thoracic cavity]
- [**Traumatic hemopneumothorax,** Fracture of one rib, Multiple fractures of ribs, Other and unspecified injuries of lung]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.3255	5,245	33	80.1	59.0	7.4	33.6

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
1.2	0.3	96.8	0.3	3.3	38.6	81.2

Cluster Q

- [Fracture of sacrum, Fracture of pubis, Fracture of lumbar vertebra, Fracture of calcaneus, Fracture of lower end of radius]
- [**Focal traumatic brain injury, Traumatic subarachnoid hemorrhage, Traumatic subdural hemorrhage**]
- [**Other disorders of fluid, electrolyte and acid-base balance,** Disorders of mineral metabolism]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.157	1,394	70	51.8	93.1	4.1	2.8

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
8.4	85.0	0.6	1.9	23.7	20.9	53.4

Cluster R

- [Traumatic pneumothorax, Multiple fractures of ribs, Other and unspecified injuries of lung]
- [**Fracture of base of skull, Focal traumatic brain injury, Traumatic subarachnoid hemorrhage, Traumatic subdural hemorrhage**]
- [**Other disorders of fluid, electrolyte and acid-base balance,** Disorders of mineral metabolism]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.137	1,177	70	52.6	90.1	5.4	4.5

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
10.7	81.1	0.9	2.4	24.4	21.9	59.7

Cluster S

- [Fracture of fourth cervical vertebra, **Fracture of fifth cervical vertebra,** Fracture of sixth cervical vertebra, Fracture of seventh cervical vertebra]
- [**Injury of colon, Injury of small intestine, Injury of other intra-abdominal organs**]
- [**Focal traumatic brain injury, Traumatic subarachnoid hemorrhage, Traumatic subdural hemorrhage**]
- [Sprain of cruciate ligament of knee, Sprain of collateral ligament of knee]
- [Fracture of metatarsal bone(s), Fracture of calcaneus, Fracture of talus, Fracture of other and unspecified tarsal bone(s)]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.122	316	40	66.1	35.1	10.4	54.4

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
80.7	4.7	1.6	10.1	12.7	1.3	89.6

Cluster T

- [Fracture of acetabulum, Subluxation and dislocation of hip, Fracture of head and neck of femur]
- [**Injury of colon, Injury of small intestine, Injury of other intra-abdominal organs**]
- [**Injury of liver and gallbladder and bile duct, Injury of spleen**]
- [Traumatic pneumothorax, Multiple fractures of ribs, Fracture of sternum]
- [Fracture of metatarsal bone(s), Fracture of calcaneus, Fracture of talus, Fracture of other and unspecified tarsal bone(s), Subluxation and dislocation of foot]

CR Score	Num Patients	Avg Age	Perc Male (%)	Mild (%)	Moderate (%)	Severe (%)
0.126	5,044	63	62.8	88.3	5.3	6.4

Mech A (%)	Mech B (%)	Mech D (%)	Mech E (%)	Mech F (%)	Mech G (%)	High Risk (%)
88.3	4.3	0.5	3.4	7.8	1.4	33.0

Figure 3: Visualization and patient characteristics of the six subgroup clusters that we discuss in Section 7. In the UMAP, we show patients in the selected subgroup in red. The remaining clusters are colored on a gradient from yellow to purple. We denote the high-risk group with the green, dashed box. High-risk injuries are bolded in the injury pattern descriptions. We highlight items of interest with the color red.

tial for a possible rescue of the patient (Kobayashi et al., 2016).

In general, we find that patients in clusters with penetrating trauma tend to be male and younger. The percentages of male patients in **Cluster O** and **Cluster P** are both around 80%, while the percentage in the training cohort is around 60% (Table 26). The average age of both clusters is 33, while the average age of the training cohort is 53.

7.3. Fall-Related Trauma Subgroups

Fall-related injury is a leading cause of death in the elderly population (CDC, 2021b), and is the most common cause of traumatic brain injury (TBI), accounting for 35% of all TBIs (Jager et al., 2000). The top clusters in the Fall category (Mech B) reflect known patterns in the medical literature. The patients tend to be older of age and female. We take **Cluster Q** and **Cluster R** as examples. The average age in the clusters is 70, versus 53 for the training cohort. The percentage male is around 51% for the two clusters, versus 60% in the training cohort.

In terms of injury patterns, **Cluster Q** and **Cluster R** include patterns of multiple fractures, combined with head injuries. Notably, both clusters also contain electrolyte imbalance disorders (too much or too little electrolytes). Though the type of electrolyte disorder is not specified, hyponatremia (low blood sodium) has been proposed to be among the factors related to elderly falls and associated with worse outcomes (Kuo et al., 2017; Rittenhouse et al., 2015). Lastly, we note that around 20% of the patients in both clusters also suffered from other blunt trauma (Mech F), which is clearly explained by falling as the primary mechanism of injury.

7.4. Motor Vehicle Accident Trauma Subgroups

Motor vehicle accidents are the second leading cause of TBI and a leading cause of death in young adults (Vella et al., 2017). Indeed, we observe that motor vehicle accidents (Mech A) clusters often include severe head injuries along with other high-risk internal injuries. **Cluster S** has a high 54% of patients with severe TBI.

Motor vehicle accidents are also the most common mechanism leading to pelvic ring and acetabulum (hip) fractures, correlated with impact direction (Dakin et al., 1999). **Cluster T** captures an injury pattern of an acetabular fracture with a femur fracture (thigh bone). In general, we find that Mech A clusters tend to have injury patterns that cover all parts of the body, from the head to the thorax to the spine to the extremities.

8. Conclusion

The identification of commonly co-occurring trauma injury patterns can aid with the earlier diagnosis of injuries. We conduct a thorough analysis of injury patterns in the TQIP cohort through learning weakly supervised representations with a β-VAE model paired with a latent classifier. We demonstrate the capability of our proposed clinical relevance score as a proxy for clinical intuition during clustering evaluation and model selection. We present a case study of selected subgroups with clinically interesting injury patterns and patient characteristics.

We envision the proposed framework can enable the more meaningful and easier extraction of injury patterns from a large, heterogeneous dataset of trauma patients. In particular, the CR score can greatly expedite the discovery process. Once the CR score is tuned to approximate clinical intuition, an ML practitioner without domain expertise can efficiently utilize this quantitative metric to select the most relevant models and extract injury patterns of interest without active feedback from clinical collaborators. There is also less burden on the clinical collaborators, as they need to review fewer model outputs, and can interact with the ML model through changes to the more interpretable CR score framework.

In this work, the validation of our developed CR score is limited by the small sample size of clinical expert collaborators we can query on an active basis. In future work, we hope to more rigorously benchmark the developed CR score with a larger group of clinical domain experts. We note that thoughtful experiment design is likely needed, as the same clinical expert may offer different absolute scales of Expert Rating over time, as we observe in Section F.1. We also hope to explore different versions of the CR score that can emphasize the discovery of novel injury patterns not previously documented in the medical literature.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu

Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Olga Andreeva, Wei Li, Wei Ding, Marieke Kuijjer, John Quackenbush, and Ping Chen. Catalysis clustering with gan by incorporating domain knowledge. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, page 1344–1352, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403187. URL https://doi.org/10.1145/3394486.3403187.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974. ISSN 0361-0926. doi: 10.1080/03610927408827101. URL http://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

CDC, 2021a. URL http://wonder.cdc.gov/ucd-icd10.html. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Mortality 1999-2020 on CDC WONDER Online Database, released in 2021. Data are from the Multiple Cause of Death Files, 1999-2020, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed on 2023-02-01.

CDC, 2021b. URL https://www.cdc.gov/falls/facts.html. Accessed on 2023-02-10.

CDC, 2023. URL https://wisqars.cdc.gov/data/non-fatal/home. Accessed on 2023-02-01.

Yale Chang, Junxiang Chen, Michael H. Cho, Peter J. Castaidi, Edwin K. Silverman, and Jennifer G. Dy. *Clustering with Domain-Specific Usefulness Scores*, pages 207–215. 2017. doi: 10.1137/1.9781611974973.24. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974973.24.

A. Chichom-Mefire, J. Palle-Ngunde, P.G. Fokam, A. Mokom-Awa, R. Njock, and M. Ngowe-Ngowe. Injury patterns in road traffic victims comparing road user categories: Analysis of 811 consecutive cases in the emergency department of a level i institution in a low-income country. *International Journal of Surgery Open*, 10:30–36, 2018. ISSN 24058572. doi: 10.1016/j.ijso.2017.11.005. URL https://linkinghub.elsevier.com/retrieve/pii/S2405857217300748.

American College of Surgeons Committee on Trauma. TQP PUF Chicago, IL, 2019 The content reproduced from the TQP PUF remains the full and exclusive copyrighted property of the American College of Surgeons. The American College of Surgeons is not responsible for any claims arising from works based on the original data, text, tables, or figures.

Leonardo Crespi, Daniele Loiacono, and Arturo Chiti. Chest x-rays image classification from $\beta-$variational autoencoders latent features. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, page 1–8, Dec 2021. doi: 10.1109/SSCI50451.2021.9660190.

Greg J Dakin, Alan W. Eberhardt, Jorge E. Alonso, James P. Stannard, and Kenneth A. Mann. Acetabular fracture patterns: associations with motor vehicle crash information. *The Journal of trauma*, 47 6:1063–71, 1999.

TraumaRegister DGU, Georg Reith, Rolf Lefering, Arasch Wafaisade, Kai O. Hensel, Thomas Paffrath, Bertil Bouillon, and Christian Probst. Injury pattern, outcome and characteristics of severely injured pedestrian. *Scandinavian Journal of Trauma,* *Resuscitation and Emergency Medicine*, 23(1):56, Dec 2015. ISSN 1757-7241. doi: 10.1186/s13049-015-0137-8. URL http://www.sjtrem.com/content/23/1/56.

Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, September 1996. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1032280214. URL https://projecteuclid.org/journals/statistical-science/volume-11/issue-3/Bootstrap-confidence-intervals/10.1214/ss/1032280214.full. Publisher: Institute of Mathematical Statistics.

Helen Fagerlind, Lara Harvey, Peter Humburg, Johan Davidsson, and Julie Brown. Identifying individual-based injury patterns in multi-trauma road users by using an association rule mining method. *Accident Analysis Prevention*, 164:106479, Jan 2022. ISSN 0001-4575. doi: 10.1016/j.aap.2021.106479. URL https://www.sciencedirect.com/science/article/pii/S0001457521005108.

Jamie L Fitch, Paul T Albini, Anish Y Patel, Matthew S Yanoff, Christian S McEvoy, Chad T Wilson, James Suliburk, Stephanie D Gordy, and S Rob Todd. Blunt versus penetrating trauma: Is there a resource intensity discrepancy? *the American Journal of Surgery*, 218.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456, 2021.

Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

Te-Cheng Hsu and Che Lin. Learning from small medical data—robust semi-supervised cancer prognosis classifier with bayesian variational autoencoder. *Bioinformatics Advances*, 3(1): vbac100, Jan 2023. ISSN 2635-0041. doi: 10.1093/bioadv/vbac100. URL https://academic.oup.com/bioinformaticsadvances/article/doi/10.1093/bioadv/vbac100/6978241.

TE Jager, HB Weiss, JH Coben, and PE Pepe. Traumatic brain injuries evaluated in u.s. emergency departments, 1992-1994. *Acad Emerg Med*, 2:134–140, 2000. doi: 10.1111/j.1553-2712.2000.tb00515.x.

Jerry J Kim, Hamid Alipour, Arthur Yule, David S Plurad, Matthew Koopmann, Brant Putnam, Christian de Virgilio, and Dennis Y Kim. Outcomes after external iliac and femoral vascular injuries. *Annals of Vascular Surgery*, 33:88–93, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Leslie M. Kobayashi, Todd W. Costantini, Michelle G. Hamel, Julie E. Dierksheide, and Raul Coimbra. Abdominal vascular trauma. *Trauma Surgery Acute Care Open*, 1(1):e000015, 2016. ISSN 2397-5776. doi: 10.1136/tsaco-2016-000015.

Robert Krajewski, Tobias Moers, Dominik Nerger, and Lutz Eckstein. Data-driven maneuver modeling using generative adversarial networks and variational autoencoders for safety validation of highly automated vehicles. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2383–2390. IEEE, 2018.

Spencer C. H. Kuo, Pao-Jen Kuo, Cheng-Shyuan Rau, Shao-Chun Wu, Shiun-Yuan Hsu, and Ching-Hua Hsieh. Hyponatremia is associated with worse outcomes from fall injuries in the elderly. *International Journal of Environmental Research and Public Health*, 14(5):460, Apr 2017. ISSN 1660-4601. doi: 10.3390/ijerph14050460.

F. Lecky, M. Woodford, A. Edwards, O. Bouamra, and T. Coats. Trauma scoring systems and databases. *British Journal of Anaesthesia*, 113(2):286–294, Aug 2014. ISSN 00070912. doi: 10.1093/bja/aeu242. URL https://linkinghub.elsevier.com/retrieve/pii/S0007091217315192.

Shengchen Li, Ke Tian, and Rui Wang. Unsupervised heart abnormality detection based on phonocardiogram analysis with beta variational auto-encoders.

In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8353–8357. IEEE, 2021.

Saaed Mehrabi, Sunghwan Sohn, Dingheng Li, Joshua J. Pankratz, Terry Therneau, Jennifer L. St. Sauver, Hongfang Liu, and Mathew Palakal. Temporal pattern and association discovery of diagnosis codes using deep learning. In *2015 International Conference on Healthcare Informatics*, page 408–416, Dallas, TX, Oct 2015. IEEE. ISBN 9781467395489. doi: 10.1109/ICHI.2015.58. URL https://ieeexplore.ieee.org/document/7349719/.

Max Metzger, Michael Howard, Lee Kellogg, and Rishi Kundi. In *2015 IEEE International Conference on Big Data (Big Data)*, page 2560–2568, Oct 2015. doi: 10.1109/BigData.2015.7364053.

R. Pfeifer and HC Pape. Missed injuries in trauma patients: A literature review. *Patient Saf Surg*, 2 (20), 2008. doi: 10.1186/1754-9493-2-20.

Katelyn J. Rittenhouse, Tuc To, Amelia Rogers, Daniel Wu, Michael Horst, Mathew Edavettal, Jo Ann Miller, and Frederick B. Rogers. Hyponatremia as a fall predictor in a geriatric trauma population. *Injury*, 46(1):119–123, Jan 2015. ISSN 00201383. doi: 10.1016/j.injury.2014.06.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0020138314003064.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov 1987. ISSN 03770427. doi: 10.1016/0377-0427(87)90125-7. URL https://linkinghub.elsevier.com/retrieve/pii/0377042787901257.

C. Satheesh, Suraj Kamal, A. Mujeeb, and M. H. Supriya. Passive sonar target classification using deep generative β-vae. *IEEE Signal Processing Letters*, 28:808–812, 2021. ISSN 1558-2361. doi: 10.1109/LSP.2021.3071255.

Nahum Sá and Itzhak Roditi. -variational autoencoder as an entanglement classifier. *Physics Letters A*, 417:127697, Nov 2021. ISSN 0375-9601. doi: 10.1016/j.physleta.2021.127697. URL https://www.sciencedirect.com/science/article/pii/S0375960121005612.

Alan H. Tyroch, Emmett L. Mcguire, Susan F. Mclean, Rosemary A. Kozar, Keith A. Gates, Krista L. Kaups, Charles Cook, Sarah M. Cowgill, John A. Griswold, Larry A. Sue, Michael L. Craun, and Jan Price. The association between chance fractures and intra-abdominal injuries revisited: A multicenter review. *The American Surgeon*, 71(5): 434–438, 2005. doi: 10.1177/000313480507100514.

MA Vella, ML Crandall, and MB Patel. Acute management of traumatic brain injury. *Surgical Clinics of North America*, 5:1015–1030, 2017. doi: 10.1016/j.suc.2017.06.003.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1073–1080, New York, NY, USA, Jun 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553511. URL https://doi.org/10.1145/1553374.1553511.

Viola Wenz, Arno Kesper, and Gabriele Taentzer. Detecting quality problems in data models by clustering heterogeneous data values. In *2021 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, page 150–159, Fukuoka, Japan, Oct 2021. IEEE. ISBN 9781665424844. doi: 10.1109/MODELS-C53483.2021.00027. URL https://ieeexplore.ieee.org/document/9643782/.

Zhongbin Xie and Shuai Ma. Dual-view variational autoencoders for semi-supervised text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, page 5306–5312, Macao, China, Aug 2019. International Joint Conferences on Artificial Intelligence Organization. ISBN 9780999241141. doi: 10.24963/ijcai.2019/737. URL https://www.ijcai.org/proceedings/2019/737.

Appendix A. Data Preprocessing of Auxiliary Features

We consider four types of auxiliary features: age, mechanism of injury, GCS, and high-risk. Age is a single feature that is normalized to the range $[0, 1]$.

Table 4: The mechanism of injury groups with their training set prevalence.

Group	Mechanism	Prevalence
A	Motor Vehicle Collision	0.308
B	Fall	0.407
C	Burn	0.014
D	Penetrating Trauma	0.089
E	Struck by Motor Vehicle	0.054
F	Other Blunt Trauma	0.234
G	Other Injury	0.137
H	Poisoning	0.002

Mechanism of Injury In the TQIP dataset, the mechanism of injury data feature ("MECHANISM") is divided into 27 categories describing the cause of the injury. Our collaborating clinician grouped these into eight larger categories (Table 4) based on the type of trauma injury the patient is expected to incur for each of the finer categories (Table 6). For instance, since both "Fire/flame" and "Hot object/substance" leads to burns, they are grouped together to form Category C (Burns).

Glasgow Coma Scale (GCS) We use the "TOTALGCS" data feature in the TQIP dataset for the raw value of the total GCS. Total GCS is the sum of the motor, verbal, and eye-opening GCS scores. Total GCS can vary on a scale from 3 to 15. We group GCS into three categories of mild, moderate, and severe head injury depending on the score (See Table 5).

Table 5: Our clinician-defined mapping from the raw total GCS values to the three broad categories.

Group	Total GCS	Prevalence
mild	14-15	0.89
moderate	9-13	0.04
severe	3-8	0.07

High-Risk Injuries We define the concept of high-risk injuries as injuries that occur at a much higher prevalence in patients who died than in patients who didn't die. With the TQIP feature "HOSPDISCHARGEDISPOSITION" = 5 (Deceased/Expired), we form a deceased subgroup of 2.6% of the patient cohort who died in the hospital. For the list of 500 selected diagnosis codes, we compute the ratio of the occurrence of the condition in the deceased group over the occurrence in the non-deceased group. We rank by this ratio and selected the top 50 conditions to be classified as high-risk injuries (Table 7). Patients with at least one high-risk injury in the set of their diagnoses are indicated with a 1. The percentage of high-risk patients in the training set is 38.6%.

Since the goal is to identify interesting trauma injury patterns during the early stages of diagnosis, we only use data features of the patient that we would have access to upon admission to the hospital. Basic information such as age and mechanism of injury is generally known before admission. GCS is evaluated at least once upon admission.

Table 6: Our clinician-defined mapping for how the TQIP "MECHANISM" data feature corresponds with the larger mechanism of injury groupings.

Mechanism of Injury (TQIP)	**Category**
1=Cut/pierce	D
2=Drowning/submersion	G
3=Fall	B
4=Fire/flame	C
5=Hot object/substance	C
6=Firearm	D
7=Machinery	G
8=MVT Occupant	A
9=MVT Motorcyclist	A
10=MVT Pedal cyclist	E
11=MVT Pedestrian	E
12=MVT Unspecified	NA
13=MVT Other	NA
14=Pedal cyclist, other	F
15=Pedestrian, other	F
16=Transport, other	F
17=Natural/environmental, Bites and stings	G
18=Natural/environmental, Other	G
19=Overexertion	G
20=Poisoning	H
21=Struck by, against	F
22=Suffocation	F
23=Other specified and classifiable	G
24=Other specified, not elsewhere classifiable	G
25=Unspecified	G
26=Adverse effects, medical care	H
27=Adverse effects, drugs	H

Table 7: List of top 50 injuries categorized to be high-risk according to the procedure outlined in Section A

ICD10	Injury Description	Ratio	Prevalence
S06.1	Traumatic cerebral edema	23.62	0.164
S06.2	Diffuse traumatic brain injury	14.82	0.057
G93	Other disorders of brain	12.26	0.014
S35.5	Injury of iliac blood vessels	9.93	0.018
S06.8	Other specified intracranial injuries	9.31	0.107
S02.9	Fracture of unspecified skull and facial bones	8.83	0.023
S35.2	Injury of celiac or mesenteric artery and branches	8.32	0.009
S26.0	Injury of heart with hemopericardium	6.97	0.013
S36.2	Injury of pancreas	6.91	0.015
S06.3	Focal traumatic brain injury	6.67	0.312
S15.0	Injury of carotid artery of neck	6.33	0.017
S02.0	Fracture of vault of skull	6.24	0.173
S25.0	Injury of thoracic aorta	5.86	0.015
S02.1	Fracture of base of skull	5.55	0.173
T21.3	Burn of third degree of trunk	5.43	0.006
S36.3	Injury of stomach	5.16	0.011
S06.6	Traumatic subarachnoid hemorrhage	4.9	0.358
S06.5	Traumatic subdural hemorrhage	4.87	0.424
S15.1	Injury of vertebral artery	4.79	0.021
S26.1	Injury of heart without hemopericardium	4.69	0.008
S36.5	Injury of colon	4.56	0.032
S27.8	Injury of other specified intrathoracic organs	4.5	0.031
S13.1	Subluxation and dislocation of cervical vertebrae	4.32	0.019
S06.4	Epidural hemorrhage	4.28	0.033
S36.4	Injury of small intestine	4.26	0.034
S14.1	Other and unspecified injuries of cervical spinal cord	4.13	0.038
S22.5	Flail chest	4.11	0.028
R40	Somnolence, stupor and coma	4.1	0.005
S36.8	Injury of other intra-abdominal organs	4.08	0.058
T22.3	Burn of third degree of shoulder and upper limb	4.07	0.005
S36.1	Injury of liver and gallbladder and bile duct	3.86	0.084
T24.3	Burn of third degree of lower limb	3.78	0.005
S27.1	Traumatic hemothorax	3.77	0.058
S37.2	Injury of bladder	3.67	0.01
S12.0	Fracture of first cervical vertebra	3.49	0.029
S12.2	Fracture of third cervical vertebra	3.39	0.014
S75.0	Injury of femoral artery	3.37	0.006
E87	Disorders of fluid, electrolyte, acid-base balance	3.33	0.006
S72.9	Unspecified fracture of femur	3.31	0.007
S37.8	Injury of other urinary and pelvic organs	3.29	0.013
S37.0	Injury of kidney	3.1	0.036
S12.3	Fracture of fourth cervical vertebra	3.07	0.017
S27.2	Traumatic hemopneumothorax	3.03	0.061
S36.0	Injury of spleen	3.01	0.068
S14.0	Concussion and edema of cervical spinal cord	2.98	0.004
S24.1	Other and unspecified injuries of thoracic spinal cord	2.87	0.009
S33.2	Dislocation of sacroiliac and sacrococcygeal joint	2.85	0.006
S02.8	Fractures of other specified skull and facial bones	2.85	0.078
S12.4	Fracture of fifth cervical vertebra	2.81	0.023
S12.1	Fracture of second cervical vertebra	2.81	0.044

ICD10	$\mathbf{w_i}$	**Description**
S04	1	Injury of cranial nerve
S06	1	Intracranial injury
S14	0.5	Injury of nerves and spinal cord at neck level
S15	0.5	Injury of blood vessels at neck level
S24	0.5	Injury of nerves and spinal cord at thorax level
S25	0.5	Injury of blood vessels of thorax
S26	1	Injury of heart
S27	1	Injury of other and unspecified intrathoracic organs
S34	0.5	Injury of lumbar and sacral spinal cord and nerves at abdomen, lower back and pelvis level
S35	0.5	Injury of blood vessels at abdomen, lower back and pelvis level
S36	1	Injury of intra-abdominal organs
S37	1	Injury of urinary and pelvic organs
S44	0.5	Injury of nerves at shoulder and upper arm level
S45	0.5	Injury of blood vessels at shoulder and upper arm level
S54	0.5	Injury of nerves at forearm level
S55	0.5	Injury of blood vessels at forearm level
S64	0.5	Injury of nerves at wrist and hand level
S65	0.5	Injury of blood vessels at wrist and hand level
S74	0.5	Injury of nerves at hip and thigh level
S75	0.5	Injury of blood vessels at hip and thigh level
S84	0.5	Injury of nerves at lower leg level
S85	0.5	Injury of blood vessels at lower leg level
S94	0.5	Injury of nerves at ankle and foot level
S95	0.5	Injury of blood vessels at ankle and foot level
XX0	-1	Superficial injury

Table 8: The weights that define the internal-external sub-metric as part of our clinical relevance scoring algorithm. Internal injuries in key regions of the head, thorax, and abdomen are given the most positive 1 weight. Injuries of blood vessels and nerves are weighed 0.5, while any superficial injury (X serves as a placeholder) is down-weighed as -1.

Appendix B. Data Preprocessing of ICD10 Codes

We visualize the training prevalence of the top 500 selected ICD10 codes in Figure 4. Since the training set prevalence drops off quickly after the top 100 codes, the cutoff of 500 most frequent codes is reasonable for our analysis.

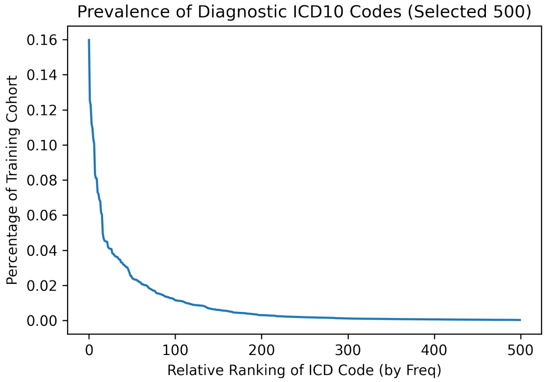


Figure 4: Plot of the selected 500 ICD10 codes ranked by prevalence in the training cohort.

Appendix C. Auxiliary Classification Task

We note that all evaluated models perform poorly at predicting moderate GCS group and the mechanism of injury groups of E, F, G, H (Table 9, Table 10, Table 11). The lower performance can be attributed to greater patient heterogeneity. We observe that the β-VAE Classifier consistently learns the separation between high-risk and low-risk groups (see Figure 7), while the β-VAE and SVD are not able to consistently do so.

Table 9: Auxiliary task performance of the SVD model.

Features	**AUC**	**F1**	**Recall**	**Prec.**
mild	0.735	0.943	0.995	0.895
moderate	0.666	0.000	0.000	0.000
severe	0.768	0.104	0.057	0.590
A	0.806	0.566	0.484	0.682
B	0.805	0.680	0.708	0.654
C	0.993	0.878	0.937	0.826
D	0.936	0.674	0.585	0.794
E	0.762	0.084	0.045	0.607
F	0.657	0.027	0.014	0.588
G	0.699	0.098	0.054	0.583
H	0.791	0.000	0.000	0.000
Risk	0.893	0.762	0.728	0.799

Table 10: Auxiliary task performance of the β-VAE model.

Features	AUC	F1	Recall	Prec.
mild	0.820	0.948	0.986	0.912
moderate	0.727	0.000	0.000	0.000
severe	0.859	0.308	0.203	0.645
A	0.802	0.556	0.481	0.662
B	0.821	0.679	0.656	0.705
C	0.993	0.882	0.918	0.849
D	0.940	0.663	0.576	0.783
E	0.764	0.061	0.032	0.608
F	0.660	0.092	0.050	0.571
G	0.694	0.059	0.032	0.581
H	0.783	0.000	0.000	0.000
Risk	0.991	0.939	0.932	0.945

Table 11: Auxiliary task performance of the β-VAE Classifier model

Features	AUC	F1	Recall	Prec.
mild	0.835	0.950	0.983	0.918
moderate	0.737	0.001	0.001	0.047
severe	0.876	0.408	0.294	0.666
A	0.830	0.608	0.542	0.692
B	0.849	0.720	0.722	0.718
C	0.994	0.895	0.941	0.854
D	0.959	0.740	0.684	0.807
E	0.787	0.142	0.082	0.537
F	0.686	0.185	0.111	0.573
G	0.725	0.200	0.121	0.581
H	0.821	0.001	0.000	0.017
Risk	1.000	1.000	1.000	1.000

Appendix D. UMAP Visualizations of Model Examples

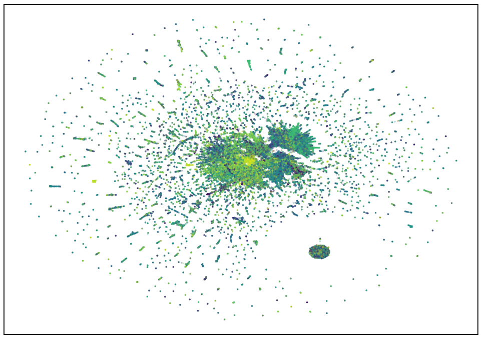


Figure 5: UMAP visualization of a typical SVD model explored in Section 5.2.

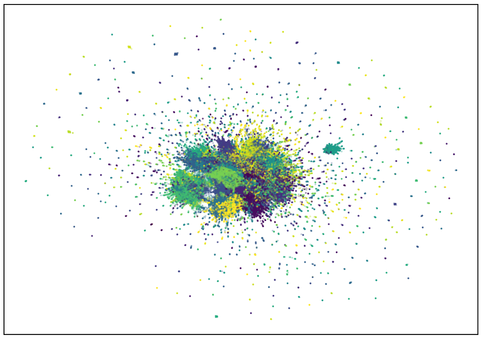


Figure 6: UMAP visualization of a typical β-VAE model explored in Section 5.2.

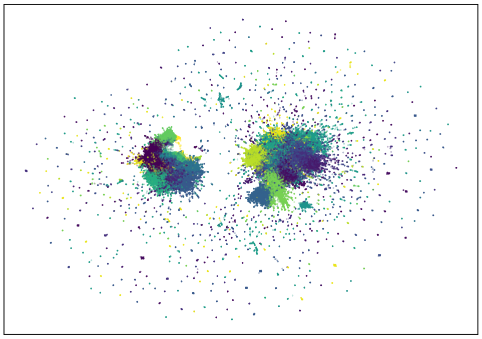


Figure 7: UMAP visualization of a typical β-VAE Classifier model explored in Section 5.2.

Appendix E. Additional Ablation and Robustness Results

E.1. Number of Clusters

Table 12: Effect of the cluster number (K) for the KMeans algorithm on the unsupervised representation performance. Metrics are averaged across 5 randomized runs of the β-VAE Classifier model.

K	CR Score	Silh. Coef.	CH Index
5	0.165	0.065	718.7
10	0.150	0.073	563.4
20	0.136	0.073	418.6
30	0.132	0.073	333.0
40	0.123	0.067	282.7
50	0.130	0.060	247.4
60	0.122	0.056	221.2
70	0.127	0.057	202.1
80	0.128	0.054	182.9
90	0.132	0.053	171.2
100	0.126	0.051	159.3

For the KMeans algorithm, we vary the number of clusters (K) and observe in Table 12 that for both the CR and the CH Index, a small number of clusters (K=5 and K=10) perform better than a larger number of clusters. The reason is that the global structure of the typical latent space of the β-VAE Classifier is divided into two main clusters corresponding to the high-risk and lower-risk groups, as we previously noted. Patients with burn injuries are also typically placed in their own cluster far from the main cohort. Thus, since there are usually 3 to 5 clouds of dispersed density, the clustering metrics are optimized for smaller cluster numbers. Similarly, the CR score is higher because if the model has only 5 clusters, then on average, most of the injury patterns in these clusters are clinically relevant. Practically, however, we are not able to extract meaningful clusters of interest with such a small number of clusters.

If we visualize the set of binary auxiliary features overlaid on the clustered UMAP visualization, we can discover some of these local clusters by eye. For instance, we see in Figure 8 that Group D (Penetrating Trauma) is primarily concentrated in two compact local areas (upper hook of the left cluster and

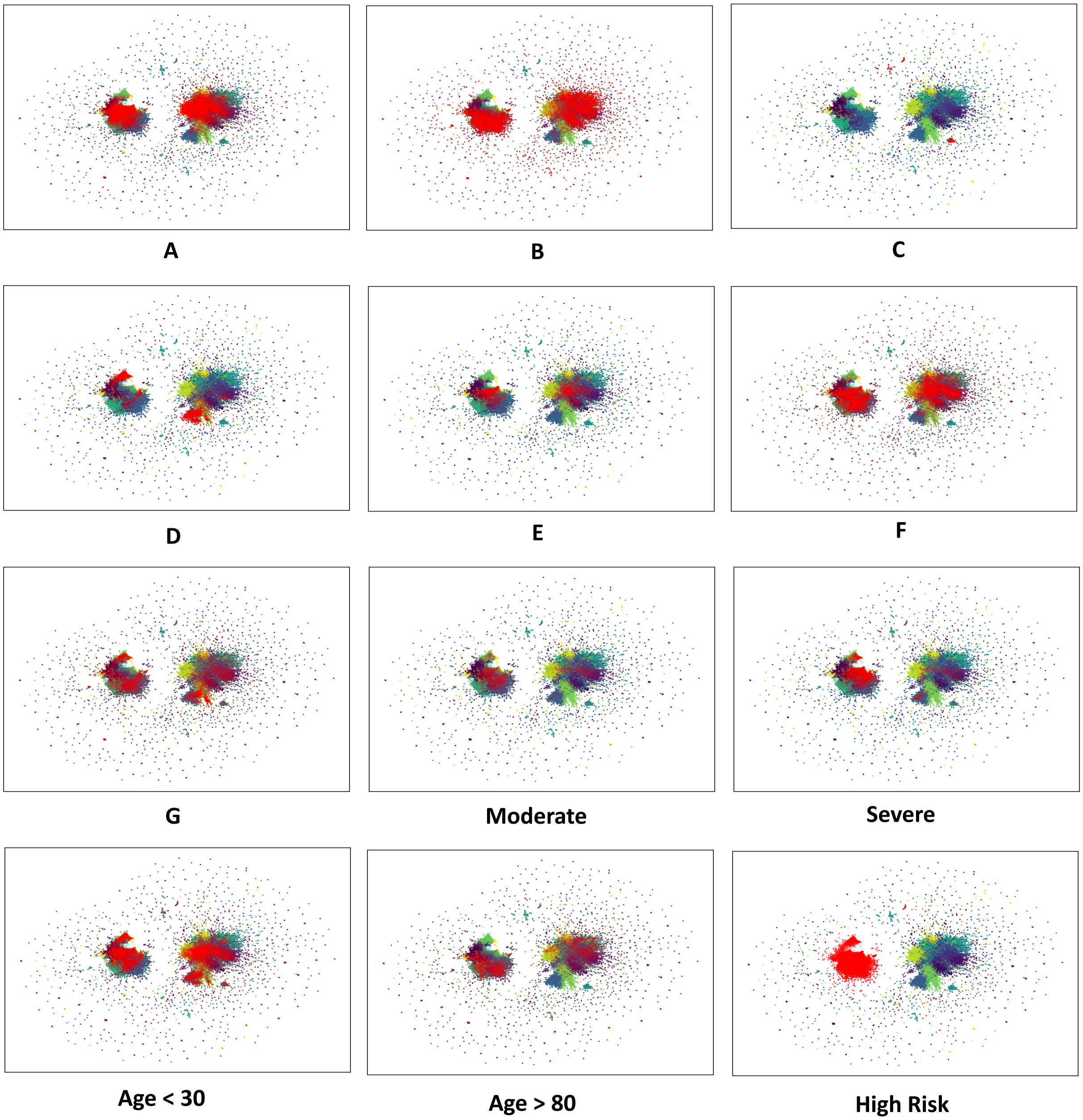


Figure 8: The patients in the positive class of each binary auxiliary signal are overlaid in red over the UMAP visualization colored by the 30 clusters. The latent embedding from the β-VAE Classifier with the highest CR score. Groups A to G can be referenced with their corresponding mechanism of injury groups. Due to space constraints, we do not include the subfigures for Group H and the Mild GCS group. Group H (Poisoning) is very sparse while the Mild group covers all cluster density.

the lower extension of the right cluster). These two groups overlap with the patients less than 30 years old subgroup, and is mostly disjoint with the patients more than 80 years old subgroup. This phenomenon is explained by the mechanism of injury, since firearms and cut/pierce are the only two valid subcategories for penetrating trauma in our data preprocessing. Thus, we confirm that the β-VAE Classifier is indeed learning informative local clusters. It's just that these local clusters may be near each other in the latent space and form larger density clouds that are easier to cluster. Based on the output of the cluster descriptions for varying K, we decide on a cluster number of K=30 for the main experiments in this work. We qualitatively feel that K=30 reasonably balances the relevance and variety of the discovered injury patterns. We note that K=30 does exhibit a reasonable CR score and clustering performance as well (Table 12).

E.2. Clustering Algorithms

Table 13: Unsupervised representation performance for different clustering algorithms averaged across 50 randomized runs of the β-VAE Classifier model.

Alg	**CR Score**	**Silh. Coef.**	**CH Index**
KMeans	0.128 (± 0.017)	**0.075 (± 0.028)**	**340.8 (± 92.7)**
BKMeans	**0.130 (± 0.042)**	0.040 (± 0.023)	274.1 (± 68.9)
Ward	0.125 (± 0.030)	0.044 (± 0.025)	281.4 (± 69.5)

Besides KMeans, we also briefly explored two other clustering algorithms. We tested agglomerative clustering with ward linkage ("Ward") and a hierarchical variant of KMeans called BisectingKMeans ("BKMeans"). In Ward agglomerative clustering, each point starts as its own cluster. During clustering, the points are linked together to minimize the sum of squared differences within all clusters. In bisecting K-Means, the clustering is hierarchical, as single clusters are successively chosen and split into new clusters (Buitinck et al., 2013).

We see in Table 13 that KMeans performs notably better the other two in terms of the clustering metrics. For the CR score, the marginal improvement of the BKMeans over the KMeans algorithm was not sufficient to justify the additional computational time, and thus we settled on using the KMeans algorithm with K=30 for our main experiments.

E.3. Consistency of Learned Representations

Table 14: The adjusted mutual information score (AMI) averaged across 50 pairs of labels with 95 confidence intervals. Model 1 and Model 2 denote which model architectures was the label sampled from.

Model 1	**Model 2**	**AMI**
SVD	SVD	0.772 (± 0.005)
BetaVAE	BetaVAE	0.449 (± 0.008)
BetaVAE Classifier	BetaVAE Classifier	0.574 (± 0.008)
SVD	BetaVAE	0.206 (± 0.004)
SVD	BetaVAE Classifier	0.317 (± 0.007)
BetaVAE	BetaVAE Classifier	0.315 (± 0.006)

To evaluate the agreement of cluster assignments of patients in the test cohort for the same and different model architectures, we computed the average adjusted mutual information score (AMI) for pairs of label sets. AMI measures agreement between two clusterings, while correcting for the effect of the agreement solely due to chance (Vinh et al., 2009). Perfect matching will have a score of 1, while a random pair will have a score of around 0. We see in Table 14 that SVD is the model that clusters the most consistently, followed by the β-VAE Classifier. In general, the cluster assignment disagrees more across different model architectures than within the same model architecture.

E.4. Ablation: Temporal Shift

To validate the claim that there are no significant temporal shifts in data (Section 3.1), we trained and evaluated the β-VAE Classifier model on patient cohorts constrained to year = {2017, 2018, 2019}. For each year, we randomly sampled 250,000 patients, and perform a 70-30 train-test split to form training cohort ($n = 175,000$) and test cohort ($n = 75,000$). Based on results in Table 15 and Table 16, we see that the average value of all metrics are similar and all confidence intervals overlap. Qualitatively, the clinical collaborators also did not detect significant differences in the discovered injury patterns across the years.

Table 15: Ablation of the year of the patient record on the auxiliary classification task for the β-VAE Classifier model (CIs over 5 randomized runs).

year	AUC	F1	Recall	Prec.
2017	0.749 (±0.018)	0.472 (±0.004)	0.446 (±0.001)	0.527 (±0.018)
2018	0.747 (±0.013)	0.473 (±0.001)	0.447 (±0.002)	0.528 (±0.008)
2019	0.745 (±0.017)	0.475 (±0.005)	0.447 (±0.004)	0.534 (±0.016)

Table 16: Ablation of the year of the patient record on the unsupervised clustering task for the β-VAE Classifier model ().

year	CR Score	Silh. Coef.	CH Index
2017	0.129 (±0.007)	0.035 (±0.038)	312.2 (±142.0)
2018	0.136 (±0.015)	0.033 (±0.031)	303.2 (±122.3)
2019	0.135 (±0.021)	0.040 (±0.021)	330.4 (±74.7)

E.5. Ablation: Model Hyperparameters

We did not find the model very sensitive to specific values of β or γ, as long as they are within the general ranges of: $\beta \in [1, 25], \gamma \in [1, 10]$.

When $\beta \geq 1$, greater latent space disentanglement is induced through upweighting the KL loss term with the isotropic Gaussian prior. We observe that auxiliary classification performance decreases with larger β (Table 17). Higher β yields better unsupervised clustering metrics, as expected through the greater disentanglement enforced in the latent space by larger β (Table 18). We use $\beta = 5$ for our main experiments, as it has the highest averaged CR score of 0.117.

In Table 19, we see that different γ performs the best for different metrics evaluating the auxiliary classification performance. As γ increases, the unsupervised clustering metrics get better (Table 20). During training, however, we note empirically that if the γ parameter is set too high (e.g. $\gamma = 25$), the KL loss term can sometimes diverge and the latent space becomes nonsensical. If we exclude $\gamma = 25$, there is no clear choice of γ given both task performances. We settle on the intuitive choice of $\gamma = 1$, which would give equal weight to the classifier loss and the reconstruction loss in the objective function.

From Table 21 and Table 22, we see that the performance on both tasks is not sensitive to the latent dimension size. We choose a latent dimension size of 64 as it has the highest CR score of 0.127.

Table 17: Ablation of the disentanglement loss hyperparameter β on the auxiliary classification task for the β-VAE model (CIs over 5 randomized runs).

β	AUC	F1	Recall	Prec.
1.0	**0.825 (±0.002)**	**0.440 (±0.004)**	**0.413 (±0.006)**	**0.609 (±0.002)**
5.0	0.821 (±0.001)	0.433 (±0.003)	0.405 (±0.004)	0.606 (±0.004)
10.0	0.819 (±0.001)	0.425 (±0.005)	0.401 (±0.005)	0.603 (±0.008)
25.0	0.815 (±0.002)	0.420 (±0.001)	0.393 (±0.002)	0.597 (±0.004)

Table 18: Ablation of the disentanglement loss hyperparameter β on the unsupervised clustering task for the β-VAE model (CIs over 5 randomized runs).

β	**CR Score**	**Silh. Coef.**	**CH Index**
1.0	0.109 (±0.025)	0.042 (±0.004)	168.8 (±15.9)
5.0	**0.117 (±0.018)**	0.043 (±0.006)	181.8 (±11.6)
10.0	0.110 (±0.011)	0.046 (±0.006)	189.8 (±18.6)
25.0	0.104 (±0.017)	**0.060 (±0.007)**	**224.2 (±14.3)**

Table 19: Ablation of the classifier loss hyperparameter γ on the auxiliary classification task for the β-VAE Classifier model (CIs over 5 randomized runs).

γ	**AUC**	**F1**	**Recall**	**Prec.**
0.1	0.841 (±0.001)	0.476 (±0.001)	0.448 (±0.003)	**0.625 (±0.001)**
1.0	**0.845 (±0.001)**	0.487 (±0.001)	0.455 (±0.001)	0.619 (±0.001)
5.0	0.839 (±0.003)	**0.490 (±0.002)**	0.459 (±0.003)	**0.625 (±0.008)**
10.0	0.833 (±0.001)	0.488 (±0.001)	0.458 (±0.001)	0.613 (±0.008)
25.0	0.830 (±0.002)	**0.490 (±0.002)**	**0.460 (±0.002)**	0.608 (±0.003)

Table 20: Ablation of the classifier loss hyperparameter γ on the auxiliary classification task for the β-VAE Classifier model (CIs over 5 randomized runs).

γ	**CR Score**	**Silh. Coef.**	**CH Index**
0.1	0.130 (±0.006)	0.053 (±0.012)	231.3 (±57.3)
1.0	0.139 (±0.005)	0.059 (±0.009)	304.2 (±37.3)
5.0	0.138 (±0.010)	0.074 (±0.014)	376.9 (±38.6)
10.0	0.139 (±0.006)	0.083 (±0.011)	417.0 (±81.0)
25.0	**0.153 (±0.013)**	**0.097 (±0.018)**	**430.1 (±78.1)**

Table 21: Ablation of the latent dimension size on the auxiliary classification task for the β-VAE Classifier model (CIs over 5 randomized runs).

latent dim	**AUC**	**F1**	**Recall**	**Prec.**
16	0.837 (±0.006)	**0.487 (±0.003)**	**0.456 (±0.002)**	0.626 (±0.033)
32	0.840 (±0.003)	**0.487 (±0.002)**	**0.456 (±0.002)**	0.613 (±0.002)
64	0.843 (±0.002)	**0.487 (±0.004)**	**0.456 (±0.003)**	0.621 (±0.012)
128	**0.844 (±0.001)**	0.485 (±0.001)	0.454 (±0.002)	0.621 (±0.007)
256	**0.844 (±0.001)**	0.485 (±0.001)	0.455 (±0.001)	**0.627 (±0.020)**

Table 22: Ablation of the latent dimension size on the unsupervised clustering task for the β-VAE Classifier model (CIs over 5 randomized runs).

latent dim	**CR Score**	**Silh. Coef.**	**CH Index**
16	0.124 (±0.013)	0.059 (±0.010)	346.1 (±83.5)
32	0.125 (±0.010)	0.054 (±0.018)	274.9 (±49.9)
64	**0.127** **(±0.025)**	0.063 (±0.009)	314.2 (±48.4)
128	0.124 (±0.024)	**0.083** **(±0.012)**	**372.5** **(±45.8)**
256	0.120 (±0.009)	0.073 (±0.036)	353.9 (±102.5)

Appendix F. Additional Analysis

F.1. Validation of Model Selection Capacity of the CR Score with More Candidate Models

After the first round of evaluation of the two models with the best CR score (relative ranking = 1) and best unsupervised clustering metrics (relative ranking = 21) in Section 5.3, we asked our clinical collaborators to perform a second round of evaluation by assigning the Expert Rating to model output. We asked the clinicians to evaluate for models at relative ranking = {10, 30, 40, 50} to approximate an interval of 10 among the candidate pool of 50 trained models.

The Pearson correlation of the CR score and Expert Rating across all six models is -0.071, which implies no correlation. However, if we only look at the four models from the second round of evaluation, we have a Pearson correlation of 0.651, which indicates a reasonable positive association. The reason for this phenomenon is that although the relative order of the Expert Rating is positively associated with the relative order indicated by the CR score within each round of evaluation, **the clinicians consistently scored models higher in the second round of evaluation**. Taking a step back, this inconsistency in the absolute of human evaluation further motivates our work that seeks to develop a consistent, empirical proxy metric for clinical intuition.

Relative Ranking	**CR Score**	**Expert Rating**	**Silh. Coef.**	**CH Index**
1	**0.168**	**1.227**	**0.037**	**233.0**
10	0.148	1.391	0.061	288.7
21	**0.144**	**1.034**	**0.092**	**497.5**
30	0.137	1.476	0.049	298.9
40	0.132	1.250	0.062	316.7
50	0.110	1.250	0.054	316.1

Table 23: Additional evaluation of latent representations of the 50 candidate β-VAE Classifier models. We include the same results of the Unsup Top Model (rank 21) and the CR Top Model (rank 1), along with models with different relevant rankings by the CR score.

F.2. Bootstrapped Performance on Auxiliary Classification Task and Unsupervised Clustering Task

As an alternate to CIs computed over 5 randomized runs, we can also compute 95% CIs through bootstrapping the test set (DiCiccio and Efron, 1996). Specifically, we bootstrapped a sample size of 100,000 patients for 50 iterations. From Table 24, we see that β-VAE Classifier is still the model with the best auxiliary classification performance. From Table 25, we see the same trends as Table 5.2. Generally, the average values and the CIs are similar to the previous result tables. Bootstrapped CIs are slightly smaller for some metrics.

Table 24: Bootstrap CIs version of Table 5.1.

Metrics	**SVD**	**BetaVAE**	**BetaVAE Classifier**
AUC	0.773 (±0.101)	0.807 (±0.102)	**0.841 (±0.094)**
F1	0.393 (±0.339)	0.402 (±0.356)	**0.485 (±0.347)**
Recall	0.374 (±0.355)	0.378 (±0.360)	**0.454 (±0.367)**
Prec.	0.553 (±0.252)	0.581 (±0.284)	**0.620 (±0.292)**

Table 25: Bootstrap CIs version of Table 5.2.

Metrics	**SVD**	**BetaVAE**	**BetaVAE Classifier**
CR Score	0.094 (±0.005)	0.133 (±0.005)	**0.136 (±0.006)**
Silh. Coef.	**0.158 (±0.010)**	0.042 (±0.001)	0.036 (±0.002)
CH Index	159.3 (±3.2)	176.1 (±1.4)	**226.4 (±4.5)**

Appendix G. Patient Characteristics

See Table 26 and Table 27.

Table 26: Baseline characteristics for the train set, TQIP 2017-2019, n = 903,267.

Attribute	**Value**
Age in years, median (IQR)	53 (33-70)
Gender, % (n)	
Female	40.1 (362,562)
Male	59.5 (540,705)
Race, % (n)	
White	74.1 (669,332)
Black or African American	14.8 (133,346)
Asian	1.9 (17,460)
American Indian	0.9 (8,203)
Native Hawaiian or other Pacific Islander	0.2 (2,232)
Unknown/other	8.0 (72,694)
Injury Severity Score (ISS), median (IQR)	9 (5-14)
ISS <= 15, % (n)	79.3 (716,065)
ISS >15, % (n)	20.7 (186,929)
Unknown	0.0 (273)
Work-related injury, % (n)	
Yes	4.4 (40,140)
No	94.5 (853,781)
Unknown	1.0 (9,346)
Inter-facility transfer, % (n)	
Yes	24.5 (221,050)
No	75.5 (682,144)
Unknown	0.0 (73)
Use of protective device (>0.1%), % (n)	
None	49.6 (448,166)
Airbag present	16.0 (144,645)
Lap belt	14.6 (132,250)
Shoulder belt	11.9 (107,345)
Helmet	6.3 (56,466)
Protective clothing (e.g. padded leather pants)	0.9 (8,462)
Protective non-clothing gear (e.g. shin guard)	0.3 (2,808)
Eye protection	0.1 (914)
Other	0.2 (1,916)
Hospital teaching status, % (n)	
University	44.0 (397,830)
Community	38.2 (345,040)
Non-teaching	17.1 (155,245)
Unkown	0.6 (5,152)
Bed size, % (n)	
>600	33.2 (299,606)
401 - 600	30.8 (277,847)
201 - 400	28.4 (256,843)
<= 200	7.6 (68,971)

IQR = interquartile range; % = percentage; n = number

Table 27: Baseline characteristics for the test set, TQIP 2017-2019, n = 259,132.

Attribute	**Value**
Age in years, median (IQR)	53 (33-70)
Gender, % (n)	
Female	40.0 (103,695)
Male	60.0 (155,437)
Race, % (n)	
White	74.0 (191,791)
Black or African American	14.8 (38,359)
Asian	1.9 (4,931)
American Indian	0.9 (2,348)
Native Hawaiian or other Pacific Islander	0.3 (657)
Unknown/other	8.1 (21046)
Injury Severity Score (ISS), median (IQR)	9 (5-14)
ISS 15, % (n)	79.5 (205,902)
ISS >15, % (n)	20.5 (53,138)
Unknown	0.0 (92)
Work-related injury, % (n)	
Yes	4.5 (11,615)
No	94.5 (244,913)
Unknown	1.0 (2,604)
Inter-facility transfer, % (n)	
Yes	24.5 (63,435)
No	75.5 (195,682)
Unknown	0.0 (15)
Use of protective device (>1%), % (n)	
None	49.5 (128,350)
Airbag present	16.0 (144,645)
Lap belt	14.6 (132,250)
Shoulder belt	11.9 (107,345)
Helmet	6.3 (56,466)
Protective clothing (e.g. padded leather pants)	0.9 (8,462)
Protective non-clothing gear (e.g. shin guard)	0.3 (2,808)
Eye protection	0.1 (914)
Other	0.2 (741)
Hospital teaching status, % (n)	
University	44.0 (113,985)
Community	38.1 (98,820)
Non-teaching	17.3 (44,903)
Unknown	0.5 (1,424)
Bed size, % (n)	
>600	33.1 (85,722)
401 - 600	31.0 (80,270)
201 - 400	28.3 (73,384)
<= 200	7.6 (19,756)

IQR = interquartile range; % = percentage; n = number