Self-Supervised Probability Imputation to Estimate the External-Natural Cause of Injury Matrix

Pirouz Naghavi NAGHAVI2@ILLINOIS.EDU
University of Illinois Urbana-Champaign, United States

Erica Naghavi EATONE3@UW.EDU
Independent Researcher, United States

Gang Wang GANGW@ILLINOIS.EDU
University of Illinois Urbana-Champaign, United States

Kanyin Liane Ong ONGL@UW.EDU
Institute for Health Metrics and Evaluation, United States

Abstract

The burden of injuries is essential to public health planning and policy making. Public health scientists rely on estimating the probability of nature-of-injuries (NI) for external causes of injuries (ECI) to calculate metrics used to describe burden of injuries globally. With more than 30 million records collected from 15 countries that include ECI with NI, in this study we develop a novel method to estimate probability of NI for ECI using self-supervised matrix imputation. We formulate learning the probability of NI for ECI for our data as a matrix imputation from noisy labels problem. Subsequently, we benchmark the collected data on 16 existing matrix imputation methods to uncover the best performing method for our data. Using self-supervision and data augmentation to curb the model's tendency to overfit to noisy labels, our matrix imputation approach improves test set RMSE by 7.36% compared to the best performing imputation model used for benchmarking. In addition, the proposed self-supervised approach reduces the Euclidean distance of NI probabilities among age groups with similar probabilities by up to 20% without impacting model performance and uses counterfactual data augmentation (CDA) to mitigate potential biases from age, sex, platform, and country income status.

Keywords: Injuries, Burden of Injuries, Matrix Imputation, Self-supervision

Data and Code Availability The code and probability matrix dataset are shared under https://github.com/ENMatrix/ENMatrix.

Institutional Review Board (IRB) This study used de-identified data, and the waiver of informed consent was reviewed and approved by the University of Washington Institutional Review Board (study number 9060).

1. Introduction

Estimating the burden of injuries is essential to public health and health policy planning, resulting in laws enacted globally to prevent fatal and non-fatal injuries. Globally, there are 4,458,185 deaths and 581,064,535 cases requiring treatment due to injuries occurring each year on average from 1990 to 2021 for all sexes and ages Institute for Health Metrics and Evaluation (IHME) (2024). Estimating the burden of fatal injuries largely relies on vital registration and verbal autopsy data, often publicly available. Estimating the burden of non-fatal injuries relies on care provider records that are often private and pertain to individual patients. In addition, injury cases are commonly recorded using either ECI, such as car accident or electrocution, or NI, such as lower limb fracture or severe burn. Despite rarely recording both NI and ECI for individual non-fatal cases, public health and health policy planning based on burden of injuries requires knowledge of both.

Disability adjusted life years (DALY) is a metric commonly used by public health scientists to estimate the burden of injuries. DALY is the sum of years of life lost (YLL) in fatal injury cases and years lived with disability (YLD) in non-fatal injury cases. Estimating YLD for a given population requires double

coded data, including NI and ECI. Using only double coded injury cases to estimate YLDs globally can produce inaccurate estimations due to limited availability of double coded data that may not represent the true distribution of injuries. Predicting missing ECI or NI for single coded data sources can improve the accuracy of the burden of non-fatal injury estimation by providing significantly more double coded data.

Public health scientists have relied on natural double coded data to learn the probability of NI given ECI in order to estimate NI cases for single coded data containing ECIs only, creating vastly more double coded data to calculate YLD Murray et al. (1996); Campos et al. (2015); Haagsma et al. (2016). NI probability given ECI is gathered in matrix format, referred to as the E-N matrix Haagsma et al. (2016). The accuracy of E-N matrix probabilities directly impacts the accuracy of calculated YLDs. In this work, we develop a novel method to estimate E-N matrices using self-supervised imputation with graph neural networks.

To estimate NI probability given ECI, we collected and assessed injury data sources from more than 15 countries, totaling over 30 million double coded injury case records. Matrix imputation is used to estimate missing values, which in our study, applies to certain NI probabilities determined from few to zero injury cases in our data.

In this work, estimating NI probabilities for ECI from our data is performed using matrix imputation from noisy labels. We benchmarked 16 existing matrix imputation methods on estimating a set of randomly selected NI probabilities from the E-N matrix. GRAPE You et al. (2020) was the best performing benchmarked model. To curb the model's tendency to overfit to noisy labels, we implemented a non-contrastive self-supervised loss for GRAPE. Our self-supervised loss reduces the Euclidean distance of NI probabilities among age groups with similar probabilities by up to 20% without increasing validation set RMSE. In our self-supervised matrix imputation approach, we learn from 7 datasets simultaneously during pre-training, 6 of which are generated using data augmentation techniques on the original data, including CDA and weighted averaging. During pre-training, we sample from a multinomial distribution generated by an autoencoder model trained by the self-supervised loss to select the dataset used in each epoch to avoid overfitting to one dataset.

Using weighted averaging data augmentation, we generate 2 datasets from the original to reduce the number of noisy labels by combining injury cases for age groups, as well as sexes and age groups, with similar probabilities. To mitigate potential bias from age, sex, platform, and country income status, we apply CDA. After pre-training, the self-supervised matrix imputation model has a similar ($\pm$5%) RMSE on the test set of CDA generated datasets and the original dataset, demonstrating CDA's effectiveness in mitigating potential bias in the dataset. Overall, our approach improves test set RMSE by 7.36% compared to GRAPE. In summary, this study makes the following contributions:

1. We collected and assessed injury data sources from more than 15 countries, totaling over 30 million double coded injury cases to estimate NI given ECI probabilities.

2. We benchmarked the dataset on 16 existing matrix imputation methods to uncover the best performing methods for our data.

3. We developed a non-contrastive self-supervised loss that reduces Euclidean distance of NI probabilities among age groups with similar probabilities by up to 20% without increasing RMSE.

4. We apply CDA to the original dataset in pre-training to mitigate potential bias from age, sex, platform, and country income status, achieving a $\pm$5% RMSE variation among CDA generated datasets and the original dataset.

5. Our self-supervised matrix imputation approach improves test set RMSE by 7.36% compared to GRAPE, the best performing matrix imputation method used during benchmarking.

2. Methodology

We describe the collected dataset and steps we took to assess the data sources to avoid introducing certain biases into the dataset. Subsequently, we describe challenges with learning from our collected dataset that includes noisy labels as a result of scarcity of injury data sources in the double coded format and splitting injury cases in the dataset into 5040 unique combinations of ECI, age group, sex, platform, and country income status for downstream tasks on public health research. Additionally, we describe our self-

supervised matrix imputation method with data augmentation, involving pre-training on the original and augmented data, and fine-tuning on the original data, to improve estimating NI probability given ECI from data that includes noisy labels.

2.1. Dataset

We curate a dataset for estimating NI probabilities from ECI by aggregating surveillance data, medical registry, hospital records, and clinical data from over 15 countries. The resulting dataset includes over 30 million double coded injury cases with ECI and NI for each case. The injury cases are sorted into different sub-groups of ECI and NI based on grouping International Classification of Diseases (ICD) codes for similar ECI and NI. Appendix E details the ICD code assignment to ECI and NI sub-groups for ICD-9 and ICD-10 coding.

We further split injury cases by age groups, including an age group for every 5 years of age up to 95 years of age with the addition of 3 groups, including less than 1 year old, 1 to 4 years old, and 95 plus years old. Injury cases are also split by sex and country income status, including high-income and not high-income, as well as platforms, including inpatient and outpatient. The age, sex, platform, and income splitting of injuries records is necessary for downstream tasks on public health research since metrics such as DALYs and YLDs are reported for different countries and demographics. The importance of platform splitting stems from DALYs and YLDs being larger for injuries involving inpatient care, since inpatient care is associated with more severe injuries. As a result, accurately estimating burden of injuries relies on knowing the care platform of injuries.

2.1.1. Data Source Assessment

Our final dataset is composed of various data sources that we examined to prevent introducing bias. Data sources that do not uniformly report injury cases, such as hospitals specializing in one area of care, are a potential source of bias. Depending on the size of the data source, there is often significant sparsity that may incorrectly appear as under-reporting or non-uniform-reporting bias. However, combining sex, income status, platform, NI, and ECI into larger theme-based groupings reduces sparsity and provides a clearer overview of the data. Figure 1 demonstrates the composition ratios of the final dataset

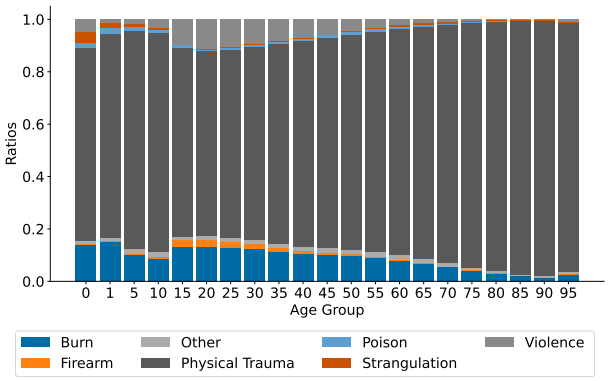

Figure 1: The composition ratios of theme-based groupings of ECI for the final dataset for all age groups after combining sexes, platforms, income statuses, and NI.

after combining ECI into larger theme-based groupings and combining sexes, platforms, income statuses, and NI. The same figure was rendered for each data source to assess the data source before adding to the dataset. Details of grouping ECI into the 7 theme-based groups is in Appendix G. Based on field experts' recommendation, including public health scientists and clinicians, data sources with skewed composition ratios were not included in the final dataset. In addition, we examined the data sources for miss-reporting biases. Using a similar approach to investigate non-uniform-reporting and under-reporting bias, we plotted counts and ratios for each of the 7 theme-based groupings of ECI. However, the NI were not combined to reveal data sources that exhibited miss-reporting bias. We considered a data source exhibiting miss-reporting bias when it frequently reported unlikely NI for ECI, and excluded them from the final dataset based on field expert recommendations. In Appendix C, we discuss other potential uses of our publicly available dataset.

2.2. Problem Statement

Estimating the probability of NI:

$$N = \{n_\alpha, n_\beta, n_\gamma, ...\}$$

given ECI:

$$E = \{e_a, e_b, e_c, ...\}$$

can be purely data driven. With enough samples collected uniformly randomly from hospitals and other

care facilities that tend and report all form of injuries, computing NI probabilities given ECI, e_a, where $e_a \in E$, is as simple as dividing the number of cases involving the NI, n_β, where $n_\beta \in N$, by the total count of all NI given ECI e_a.

$$P(n_\beta|e_a) = \frac{count(n_\beta|e_a)}{\sum_{i \in N} count(n_i|e_a)}$$

2.2.1. Limitations of Data-Driven Approach

The difficulty with the purely data driven approach lies in having to collect double coded data that contains the ECI and NI. However, double coded data is limited since most health records either include the ECI or NI. The next challenge arises when probabilities need to be estimated for different age groups, sexes, platforms (e.g. inpatient and outpatient), and country income status (e.g. high-income and not high-income). Providing NI probabilities for each 5040 unique combination of age group, sex, platform, country income status, and ECI involves separating the collected injury cases. The separation of injury cases for the 5040 unique combinations results in missing values in the dataset. After splitting the data, the extent of missing values for each unique combination involves 807 or 16% of combinations having no injury cases assigned to them, 1362 or 27% of combinations having less than 5 cases assigned to them, and 2209 or 44% of combinations having assigned cases less than 52, the total number of NI groups. The combinations with zero cases assigned to them are especially problematic using the purely data driven approach due to the inability to assign any probability to the NI given an ECI. Thus, using the purely data driven approach would be incomplete and potentially inaccurate for a considerable number of unique combinations.

2.2.2. Imputation

A common approach for estimating missing values in a dataset is feature or matrix imputation Dempster et al. (1977); Burgette and Reiter (2010); Kim et al. (2004); Cai et al. (2010). Using imputation, we can assign probabilities to NI that have zero or a small number of cases assigned to them. It is possible to set an injuries case threshold for missing data and use imputation methods to estimate missing values, however the threshold would be arbitrarily selected and may incorrectly categorize true values as missing. There are many cases where zero or low case count is the true value, such as breaking a lower limb, a NI, as a result of inhaling poison gas, an ECI. To truly identify missing values, we have to rely on expert intervention.

Missing Value Identification In our dataset, missing values cannot be feasibly distinguished from true zeros or low case counts. The dataset has 66% of NI probabilities over all combinations set to zero due to not having any injury cases assigned to them. An even larger percentage of NI have 5 or less cases assigned to them. As a result, distinguishing true zeros or low counts from missing values would require unfeasible hours of field expert intervention due to the number of cases. Thus, using traditional imputation to estimate missing values is not the optimal approach since identifying missing values in the dataset is unfeasible.

2.3. Self-Supervised Matrix Imputation

Using a randomly selected portion of the dataset with the probabilities from the purely data driven approach on NI as the training set, we can rely on feature or matrix imputation models to impute the probabilities of other parts of the dataset, used as the test and validation sets. We can estimate the probabilities for the entire dataset by repeating this approach on multiple samples of the dataset without replacement and aggregating the probabilities of the test sets. The disadvantage of this approach lies in imputation models being commonly implemented using supervised learning You et al. (2020); Burgette and Reiter (2010), and in supervised learning, noisy samples (e.g. samples with missing data) can cause the model to overfit to those noisy samples. However, past works demonstrated that self-supervised learning can improve robustness and uncertainty Hendrycks et al. (2019) as well as prevent models from overfitting to noisy instances in the dataset Tu et al. (2023).

2.3.1. Imputation Benchmarking

We first benchmark our dataset on existing feature and matrix imputation methods since our dataset has not been previously benchmarked on such methods and these methods can make assumptions about the input data that may not be appropriate for our dataset. During this step, we used a variety of matrix and feature imputation approaches, including joint modeling with expectation-maximization (EM), multivariate imputation by chained equations (MICE), k-

nearest neighbors (KNN), and matrix completion, as well as more recent approaches involving deep generative models and graph neural networks (GNN). Subsequently, we further enhanced model performance by implementing self-supervised learning in tandem with the best performing imputation approach found during benchmarking.

2.3.2. Self-Supervised Loss

We train our model to predict similar NI probabilities for scenarios with slight variations based on observations in the data. Specifically, we observed that among certain age groups, NI probabilities are very similar for ECI. An example of this scenario is a 30 year old male from a high-income country admitted for over 24 hours to a care provider for a car accident having similar NI probabilities as 25, 35, and 40 year old males in a high-income country admitted to a care provider for inpatient care. Self-supervised loss can reduce the difference in predicted NI probabilities among age groups that should have similar probabilities. After confirming our intuition with field experts, including clinicians and public health experts, we created larger age groups with similar NI probability for ECI (see Appendix F). Using the larger age groups, we developed a loss function that aligns with non-contrastive self-supervised loss discussed in past works Balestriero and LeCun (2022). Our self-supervised loss function takes the average Euclidean distance of NI probabilities for ages within the larger age groups with similar probabilities. The impact of self-supervised loss on reducing the difference in predicted NI probabilities among age groups with similar probabilities is evaluated in Section 3.2.

Self-Supervised Loss Formalization The self-supervised loss $\mathcal{L}_D$, is formalized below with NI probabilities vector

$$\boldsymbol{p_{n,e_i,s_j,a_k,l_t,m_w}} = \langle P(n_\alpha|e_i, s_j, a_k, l_t, m_w), P(n_\beta|e_i, s_j, a_k, l_t, m_w), ...\rangle$$

for NI n_α, $n_\beta \in N$. The Euclidean distance between NI probability vectors $\boldsymbol{p_{n,e_i,s_j,a_k,l_t,m_w}}$ and $\boldsymbol{p_{n,e_i,s_j,a_u,l_t,m_w}}$, denoted as $d_{e_i,s_j,l_t,m_w}(a_k, a_u)$ is

$$d_{e_i,s_j,l_t,m_w}(a_k, a_u) = \left\| \boldsymbol{p_{n,e_i,s_j,a_k,l_t,m_w}} - \boldsymbol{p_{n,e_i,s_j,a_u,l_t,m_w}} \right\|_2$$

for age groups a_k, $a_u \in A_b$ where A_b is the larger age grouping, such that $A_b \in A$ where A is the set of larger age groupings. In the loss formalization below, the ECI is $e_i \in E$, sex is $s_j \in S$ where $S = \{\text{Male, Female}\}$, platform is $l_t \in L$ where $L = \{\text{Inpatient, Outpatient}\}$, and country income status is $m_w \in M$ where $M = \{\text{High-income, Not high-income}\}$.

$$\mathcal{L}_D = \sum_{\substack{A_b \in A,\ e_i \in E,\\ s_j \in S,\ l_t \in L,\\ m_w \in M}} \frac{\frac{1}{|A_b|} \sum_{a_k, a_u \in A_b} d_{e_i,s_j,l_t,m_w}(a_k, a_u)}{|A| + |E| + |S| + |L| + |M|}$$

Combined Loss As stated in Section 2.3, we include the self-supervised loss in the objective function during training to help prevent the model from overfitting to noisy samples. Thus, the self-supervised loss is applied in tandem with the existing loss function of the imputation task, such as mean squared error (MSE) loss. Combining the self-supervised loss $\mathcal{L}_D$ with the imputation loss $\mathcal{L}_{MSE}$ aligns with our objective for the self-supervised matrix imputation training, which is learning to impute missing values in a dataset containing noisy labels without overfitting to those noisy labels. When combining two losses, the magnitude of the individual loss values can impact the training. As a result, a tunable scale factor θ_D is used to scale the two losses. The final combined loss is formalized below.

$$\mathcal{L} = \mathcal{L}_{MSE} + \theta_D \times \mathcal{L}_D$$

Excluding Loss of Known Values In our dataset, the age group, sex, platform, country income status, and ECI are always included, however NI probabilities are imputed for missing values. During training, the imputation loss is commonly computed on all input values, including values that are always known in our dataset. Thus, masking out the known values from the loss calculation during training can potentially improve model performance. Known value masking is beneficial because in certain scenarios, predicting sex or age group using other values is incorrect, given that sex or age group may not be determining factors.

2.3.3. Data Augmentation

In addition to self-supervision, we rely on counterfactual and weighted averaging data augmentation techniques to prevent our imputation model from overfitting to noisy labels. Data augmentation has been used in past works to improve general model performance Wang et al. (2019), imputation performance

Wang et al. (2021), robustness Rebuffi et al. (2021), and model performance when learning from noisy labels in certain applications Song et al. (2024). As previously stated in Section 2.3, our dataset contains noisy NI probability labels that may lead our imputation model to overfit to noisy samples. Variations in number of injury cases by age group, sex, platform, and country income status can introduce biases that result in noisy NI probability for labels driven from limited data that may not represent the true distribution.

Counterfactual CDA has been used in the past to mitigate biases in datasets in a variety of applications Dinan et al. (2020); Dash et al. (2022). Given that NI probability labels are driven by the number of cases assigned to them, data availability variation by age group, sex, platform, and country income status can introduce biases into the dataset that make NI probability labels noisy. To mitigate potential biases in our dataset that contribute to noisy labels, we implemented CDA for sex, country income status, and platform, such that we assign the opposite value, creating three additional datasets from the original data. In addition, we implemented CDA for age group, where we randomly assign age groups within larger age groups that have similar probability of NI for ECI (see Appendix F), creating an additional dataset from the original dataset.

Weighted Average Splitting injury cases by age groups, sex, platform, and country income status to predict the NI probabilities given the ECI results in splitting the data into 5040 unique combinations that may sometimes have very few injury cases that may not represent the true distribution of the NI for the combination due to limited data. However, our consultation with field experts confirmed that within larger age groups, probability of NI for ECI are similar (see Appendix F). Additionally, sex is not a determining factor for NI probabilities given the ECI with a few exceptions. Thus, we take the average of the probabilities for the combinations within the larger age groups that have similar probability of NI for ECI, however we weight the average by the number of zeros within the probability vector. Subsequently, we assign the weighted average vector to all the combinations within the larger age group, creating another dataset with weighted averages of the larger age groups. In addition, we combine sexes within the larger age groups to calculate the weighted average, creating an additional dataset with weighted averages of the larger age groups with the sexes combined.

2.3.4. Combined Approach

With the aid of the data augmentation techniques, we generated 6 new datasets from the original dataset, 4 of which are generated using CDA and the other 2 are generated using weighted averaging. Learning from 7 datasets during training is non-trivial given the model can overfit to one dataset and underfit to others. To avoid unevenly learning from the datasets, we rely on a model to generate probabilities used in a multinomial distribution to sample and select a dataset to learn from in a given epoch. The dataset selection probability generator model is an autoencoder used in past works for generative tasks Vértes and Sahani (2018) (see Appendix D). The objective of the dataset selection probability generator model is to minimize the loss on all datasets. As a result, we take the softmax of the generated probabilities for dataset selection:

$$\boldsymbol{p_d} = softmax(\langle p_{d1}, p_{d2}, ..., p_{d7} \rangle)$$

Subsequently, we take the one norm of the generated dataset selection probabilities vector multiplied by the loss used for the self-supervised imputation model

$$\mathcal{L} \times \|\boldsymbol{p_d}\|_1$$

that results in the combined loss

$$\mathcal{L}_{comb} = \mathcal{L} \times \|\boldsymbol{p_d}\|_1 = \mathcal{L}$$

In this implementation, the loss of the dataset probability generator model is the same value as the imputation model, which includes the self-supervised loss. However, the loss of the imputation model drives the optimization of all models involved in the training process, including the dataset selection probability generator model.

Pre-Training and Fine-Tuning Learning from the 7 datasets is done in the pre-training stage of our approach described in Section 2.3.4, during which the model learns from all 7 datasets to mitigate the biases present in our dataset and improve model robustness against noisy labels. In the fine-tuning stage, the model is fine-tuned to the original dataset to improve the model's performance on the original data. Although the fine-tuning step can potentially add biases back into the model predictions and result in

the model's reduced robustness against noisy labels, fine-tuning can be controlled by adjusting the initial learning rate, making the fine-tuning step more or less aggressive Wortsman et al. (2022).

3. Results

In this section, we implemented and tested the methodology defined in Section 2 on our dataset. Initially, we evaluate a range of existing imputation approaches to benchmark our dataset (see Section 3.1). Subsequently, we implement and evaluate our approach (see Section 3.2).

3.1. Imputation Benchmarking

Imputation benchmarking helps us uncover the most appropriate imputation approach for our dataset. Imputation of missing values in a dataset can be done using different approaches, including Expectation-Maximization (EM) Dempster et al. (1977), multivariate imputation by chained equations (MICE) Burgette and Reiter (2010), k-nearest neighbors (KNN) Kim et al. (2004), and matrix completion Cai et al. (2010). In recent years, deep learning has been used for matrix imputation Gondara and Wang (2018), including generative models, such as GAIN Yoon et al. (2018), and graph neural networks, such as GRAPE You et al. (2020). In recent works, OT-based imputer Muzellec et al. (2020), MIRACLE Kyono et al. (2021), TDM Zhao et al. (2023), and IGRM Zhong et al. (2023) have made improvements to imputing missing values in datasets using various techniques. The imputation benchmarking of our dataset includes the aforementioned approaches with the addition of a few other imputation techniques implemented in fancyimpute Rubinsteyn and Feldman.

Performance Evaluation To compare the performance of the imputation models, we randomly select 12.5% of the NI probabilities across the 5040 unique combinations in the dataset for the test set. We also randomly select 12.5% of the probabilities for validation, using the other 75% of the probabilities for training with the addition of known columns in the tabulated dataset, including ECI, age group, sex, platform, and country income status. The imputation methods were compared using test set RMSE. Hyper-parameters were manually tuned for the imputation methods by relying on validation set RMSE (see Appendix H).

Table 1: Matrix imputation benchmarking results

Method	**RMSE**	**MAE**
Random	0.1025	0.0482
Mean	0.0746	0.0231
Dirichlet RG (current)	0.0575	0.0154
KNN	0.0608	0.0163
MICE	0.0776	0.029
EM	0.0981	0.043
SoftImpute	0.0757	0.021
IterativeSVD	0.0732	0.0213
Matrix Factorization	0.0791	0.0316
GAIN	0.0795	0.018
Round Robin MLP	0.0824	0.0377
MIRACLE (Mean)	0.0688	0.1732
MIRACLE (GRAPE)	0.0692	0.1725
OT Imputer	0.0774	0.0286
TDM	0.0678	**0.0123**
IGRM	0.063	0.0185
GRAPE	**0.053**	0.0156
Self-Supervised GRAPE	**0.0491**	**0.0121**

Benchmarking Results The RMSE and MAE of different methods on the test set is provided in Table 1. The best performing methods on our dataset are GRAPE You et al. (2020) with RMSE of 0.053 and TDM Zhao et al. (2023) with MAE of 0.0123. The worst performing method is Random with RMSE of 0.1025 that involves assigning values between 0 and 1 to the missing probability of NI for ECI, followed by scaling all probabilities of all unique combinations to sum to one. The method currently used by public health scientists relies on Dirichlet regressions Maier (2014) with 0.0575 RMSE Haagsma et al. (2016). We selected GRAPE as the model to implement our self-supervised matrix imputation approach since model performance with regards to RMSE is more critical due to the higher influence of larger errors in RMSE compared to MAE.

3.2. Self-Supervised Matrix Imputation

After developing the self-supervised matrix imputation described in Section 2.3 with GRAPE You et al. (2020), we pre-trained a model on our dataset with our combined approach involving self-supervised loss and data augmentation generated datasets. We trained the model for 50,100 epochs, closely matching the benchmarking hyper-parameters with the exception of the learning rate scheduler (see Appendix

H). The pre-trained model obtained 0.0503 RMSE on the test set of the original dataset. At the end of pre-training, the model performed similarly on all CDA generated datasets with test set RMSE of 0.050, demonstrating the bias mitigation effectiveness of CDA. However, after fine-tuning, involving 20,000 epochs with an initial learning rate set to 28% of the initial learning rate of the pre-training, we further improved the model performance to 0.0491 RMSE and 0.0121 MAE on the test set of the original dataset (see Appendix H).

Self-Supervised Loss The self-supervised loss implemented on GRAPE takes over 10 times longer for an epoch, mostly spent on backwards propagation, compared to MSE loss used for feature imputation tasks on the original GRAPE. As a result, using the self-supervised loss for 50,000 epochs would significantly slow down the training process. However, the self-supervised loss is very effective in reducing the distance between the combinations within the larger age grouping that should have similar NI probabilities for ECI. To benefit from the self-supervised loss in practice, we introduce a cold-start-delay that allows the model to train for a number of epochs using the MSE loss before switching the objective function to the combined loss described in Section 2.3. In our implementation, we train the model for 50,000 epochs using the MSE loss before switching to the combined loss for the remaining 100 epochs. Using the combined loss for the remaining 100 epochs reduces the distance between the combinations within the larger age groups by up to 20% using the conservative scaling factor θ_D of 0.005 on the distance loss that does not increase validation set RMSE (see Appendix B).

4. Related work

Burden of injuries is commonly estimated with the method proposed by Murray et al. (1996), which uses Dirichlet regressions Maier (2014) to estimate the E-N matrix Campos et al. (2015); Haagsma et al. (2016); Ferrari et al. (2024). In this work, we formulated estimating the probability of NI for ECI as a matrix imputation from noisy labels problem for our dataset. Thus, in this section, we focus on related works involving matrix imputation.

Matrix Imputation Matrix imputation is a highly discussed topic, proposing various methods to impute missing features and samples in datasets. One such approach is statistical methods, which include EM Dempster et al. (1977); Ghahramani and Jordan (1993); García-Laencina et al. (2010); Honaker et al. (2011); Zhou et al. (2020), MICE Raghunathan et al. (2001); Van Buuren (2007); Burgette and Reiter (2010); Van Buuren and Groothuis-Oudshoorn (2011); Stekhoven and Bühlmann (2012); Laqueur et al. (2022), KNN Kim et al. (2004); Troyanskaya et al. (2001a); Keerin and Boongoen (2021), and matrix completion Troyanskaya et al. (2001b); Srebro et al. (2004); Cai et al. (2008); Candès and Recht (2008); Mazumder et al. (2010); Hastie et al. (2015); Gui et al. (2023).

Another approach to matrix imputation involves deep learning Gondara and Wang (2018); Nazabal et al. (2020); Vincent et al. (2008); Ivanov et al. (2019); Mattei and Frellsen (2019); Gong et al. (2021); Peis et al. (2022), including generative adversarial networks (GAN), such as GAIN Yoon et al. (2018), GAMIN Yoon and Sull (2020), MISGAN Li et al. (2019), MI-GAN Dai et al. (2021), and FragmGAN Fang and Bao (2024); deep flow models, such as Mcflow Richardson et al. (2020), EMFlow Ma and Ghosh (2021), and AST-CMCN Wang et al. (2022); and graph neural networks, such as GRAPE You et al. (2020), GC-MC Berg et al. (2017), IGMC Zhang and Chen (2019), GINN Spinelli et al. (2020), GEDI Chen et al. (2023), VISL Morales-Alvarez et al. (2022), MEGAE Gao et al. (2023), IGRM Zhong et al. (2023), and other GNN based works Vinas et al. (2021). In recent works, OT-based imputer Muzellec et al. (2020), MIRACLE Kyono et al. (2021), and TDM Zhao et al. (2023) proposed refinements to improve matrix imputation. Self-supervision can also improve matrix imputation, as demonstrated in recent works Liu et al. (2024). However, self-supervision for matrix imputation to learn from noisy labels is a novel method introduced in our study.

Although many approaches have relied on generative models, such as GAIN Yoon et al. (2018), HH-VAEM Peis et al. (2022), and SimpDM Liu et al. (2024), they rely on single observation as input to impute the missing values, which is difficult to achieve optimal results on a dataset with a considerable portion of its labels being noisy, given that 16% of our observations have 0 for all NI probabilities. Alternatively, approaches such as GRAPE You et al. (2020) rely on multiple observations as input to impute the missing values, in addition to mean aggregation function that can also potentially improve the model's robustness against overfitting to noisy labels. However, these model attributes of GRAPE are not spe-

cific to developing robustness against noisy labels. Thus, we make new additions to GRAPE, including self-supervised loss, weighted averaging data augmentation, and bias mitigation techniques to further improve the model's performance in the presence of noisy labels.

5. Discussion

Compared to the widely adopted method, using Dirichlet regression, our self-supervised matrix imputation approach improves test set RMSE by 14.6% under similar testing conditions. The accuracy of E-N matrix probabilities directly impacts the accuracy of calculated YLDs used to estimate the burden of injuries globally, essential to public health and health policy planning. To highlight the significance of our work, we conduct case studies focused on motorcycle accidents, interpersonal violence, car accidents, falls, and firearms. Details of the case studies are in Appendix A. In summary, our approach improves average RMSE by 91.7% for motorcycle accidents, 70.4% for interpersonal violence, 79.4% for car accidents, 88.3% for falls, and 82.9% for firearms compared to the existing Dirichlet regression approach.

5.1. Limitations

Our self-supervised matrix imputation approach mitigates the effect of noisy labels in our dataset, improving model performance in the presence of noise. Given the scope of our evaluation focuses on our dataset, the effectiveness of our approach in improving model robustness to overfitting to noisy labels has not been tested on other datasets. In addition, our approach improves the performance of the imputation model in the presence of noisy labels when removal or correction of noisy labels is not a feasible option. Thus, our approach is not intended to be a substitute for perfectly labeled data. Another consideration with regards to our work is that our method and the E-N matrix in general are intended for estimating public health metrics, such as YLDs of injuries. Since our approach is based on correlations in our dataset and does not make claims about causal analysis, it is not meant for any scope beyond predicting estimates for various public health metrics for different demographics. In terms of technical constraints, if a new ECI, NI, or age grouping is added to our dataset, our model needs to be retrained to accurately make predictions on the newly added data. In addition to the added time for retraining, performing the computation and gradient updates for the combined loss, in epochs where combined loss is used, is approximately 10 times slower for an epoch than MSE loss used in GRAPE.

6. Conclusion

In this work, we develop a novel approach to estimate NI probabilities given ECI, essential in estimating burden of injuries for public health and health policy planning. We benchmark our dataset of 30 million injury cases on 16 existing matrix imputation models. Additionally, we develop and test our self-supervised approach to curb the model's tendency to overfit to noisy labels. Our approach improves test set RMSE by 7.36% compared to the best performing model in benchmarking, reducing Euclidean distance of NI probabilities by up to 20% among age groups with similar probabilities and mitigating potential bias in our data using CDA.

7. Acknowledgments

We thank our reviewers for their highly constructive comments. This work is supported by the Bill & Melinda Gates Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26671–26685. Curran Associates, Inc., 2022.

Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

Dhanajit Brahma. Generative adversarial imputation networks (gain) pytorch implementation, 2024. URL https://github.com/dhanajitb/GAIN-Pytorch.

Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9): 1070–1076, 2010.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20:1956–1982, 2008.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Mônica Rodrigues Campos, Vanessa dos Reis von Doellinger, Luiz Villarinho Pereira Mendes, Maria de Fatima dos Santos Costa, Thiago Góes Pimentel, and Joyce Mendes de Andrade Schramm. Morbidity and mortality associated with injuries: results of the global burden of disease study in brazil, 2008. *Cadernos de Saúde Pública*, 31(1): 121–136, 2015.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2008.

Katrina Chen, Xiuqin Liang, Zheng Ma, and Zhibin Zhang. Gedi: A graph-based end-to-end data imputation framework. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 723–730. IEEE, 2023.

Zongyu Dai, Zhiqi Bu, and Qi Long. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 791–798. IEEE, 2021.

Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 915–924, January 2022.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, 2020.

Rebecca Y. Du, Melissa A. LoPresti, Roxanna M. García, and Sandi Lam. Primary prevention of road traffic accident–related traumatic brain injuries in younger populations: a systematic review of helmet legislation. *Journal of Neurosurgery: Pediatrics PED*, 25(4):361 – 374, 2020. doi: 10.3171/2019.10.PEDS19377. URL https://thejns.org/pediatrics/view/journals/j-neurosurg-pediatr/25/4/article-p361.xml.

Fang Fang and Shenliao Bao. Fragmgan: generative adversarial nets for fragmentary data imputation and prediction. *Statistical Theory and Related Fields*, 8(1):15–28, 2024.

Alize J Ferrari, Damian Francesco Santomauro, Amirali Aali, Yohannes Habtegiorgis Abate, Cristiana Abbafati, Hedayat Abbastabar, Samar Abd ElHafeez, Michael Abdelmasseh, Sherief Abd-Elsalam, Arash Abdollahi, et al. Global incidence, prevalence, years lived with disability (ylds), disability-adjusted life-years (dalys), and healthy life expectancy (hale) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet*, 403(10440):2133–2161, 2024.

Ziqi Gao, Yifan Niu, Jiashun Cheng, Jianheng Tang, Lanqing Li, Tingyang Xu, Peilin Zhao, Fugee Tsung, and Jia Li. Handling missing data via max-entropy regularized graph autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7651–7659, 2023.

Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19:263–282, 2010.

Zoubin Ghahramani and Michael Jordan. Supervised learning from incomplete data via an em approach. *Advances in neural information processing systems*, 6, 1993.

Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Ad-*

vances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22, pages 260–272. Springer, 2018.

Yu Gong, Hossein Hajimirsadeghi, Jiawei He, Thibaut Durand, and Greg Mori. Variational selective autoencoder: Learning from partially-observed heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 2377–2385. PMLR, 2021.

Brianna Greenberg, Alexandria Bennett, Asad Naveed, Raluca Petrut, Sabrina M Wang, Niyati Vyas, Amir Bachari, Shawn Khan, Tea Christine Sue, Nicole Dryburgh, et al. How firearm legislation impacts firearm mortality internationally: A scoping review. *Health Policy OPEN*, page 100127, 2024.

Yu Gui, Rina Barber, and Cong Ma. Conformalized matrix completion. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4820–4844. Curran Associates, Inc., 2023.

Juanita A Haagsma, Nicholas Graetz, Ian Bolliger, Mohsen Naghavi, Hideki Higashi, Erin C Mullany, Semaw Ferede Abera, Jerry Puthenpurakal Abraham, Koranteng Adofo, Ubai Alsharif, et al. The global burden of injury: incidence, mortality, disability-adjusted life years and time trends from the global burden of disease study 2013. *Injury prevention*, 22(1):3–18, 2016.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

James Honaker, Gary King, and Matthew Blackwell. Amelia ii: A program for missing data. *Journal of statistical software*, 45:1–47, 2011.

Institute for Health Metrics and Evaluation (IHME). Gbd compare, 2024. URL https://vizhub.healthdata.org/gbd-compare/.

O Ivanov, M Figurnov, and D Vetrov. Variational autoencoder with arbitrary conditioning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Trudy A Karlson. Injury control and public policy. *Critical Reviews in Environmental Science and Technology*, 22(3-4):195–241, 1992.

Phimmarin Keerin and Tossapon Boongoen. Improved knn imputation for missing values in gene expression data. *Computers, Materials and Continua*, 70(2):4009–4025, 2021.

Ki-Yeol Kim, Byoung-Jin Kim, and Gwan-Su Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5:1–9, 2004.

Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23806–23817. Curran Associates, Inc., 2021.

Hannah S Laqueur, Aaron B Shev, and Rose MC Kagawa. Supermice: an ensemble machine learning approach to multiple imputation by chained equations. *American journal of epidemiology*, 191(3): 516–525, 2022.

Elton Law. impyute, 2017. URL https://impyute.readthedocs.io/en/master/.

Lois K Lee, Michael C Monuteaux, Lindsey C Burghardt, Eric W Fleegler, Lise E Nigrovic, William P Meehan, Sara A Schutzman, and Rebekah Mannix. Motor vehicle crash fatalities in states with primary versus secondary seat belt laws: a time-series analysis. *Annals of internal medicine*, 163(3):184–190, 2015.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Yixin Liu, Thalaiyasingam Ajanthan, Hisham Husain, and Vu Nguyen. Self-supervision improves diffusion models for tabular data imputation. *arXiv preprint arXiv:2407.18013*, 2024.

Qi Ma and Sujit K Ghosh. Emflow: Data imputation in latent space via em and deep flow models. *arXiv preprint arXiv:2106.04804*, 2021.

Marco Maier. Dirichletreg: Dirichlet regression for compositional data in r. 2014.

Aurélie Martin, Emmanuel Lagarde, and L Rachid Salmi. Burden of road traffic injuries related to delays in implementing safety belt laws in low-and lower-middle-income countries. *Traffic injury prevention*, 19(sup1):S1–S6, 2018.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010.

James A Mercy, Susan D Hillis, Alexander Butchart, Mark A Bellis, Catherine L Ward, Xiangming Fang, and Mark L Rosenberg. Interpersonal violence: global impact and paths to prevention. *Injury prevention and environmental health. 3rd edition*, 2017.

Pablo Morales-Alvarez, Wenbo Gong, Angus Lamb, Simon Woodhead, Simon Peyton Jones, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Simultaneous missing value imputation and structure learning with groups. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20011–20024. Curran Associates, Inc., 2022.

Christopher JL Murray, Alan D Lopez, World Health Organization, et al. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary*. World Health Organization, 1996.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7130–7140. PMLR, 13–18 Jul 2020.

Aleksandra H Natora, Jennifer Oxley, Linda Barclay, Kelvin Taylor, Bruce Bolam, and Terry P Haines. Improving policy for the prevention of falls among community-dwelling older people—a scoping review and quality assessment of international national and state level public policies. *International journal of public health*, 67:1604604, 2022.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

Jay Patel, Katherine Leach-Kemon, Gwenetta Curry, Mohsen Naghavi, and Devi Sridhar. Firearm injury—a preventable public health issue. *The Lancet Public Health*, 7(11):e976–e982, 2022.

Ignacio Peis, Chao Ma, and José Miguel Hernández-Lobato. Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo. *Advances in Neural Information Processing Systems*, 35:35839–35851, 2022.

Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, Peter Solenberger, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Trevor W Richardson, Wencheng Wu, Lei Lin, Beilei Xu, and Edgar A Bernal. Mcflow: Monte carlo flow models for data imputation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14205–14214, 2020.

Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python. URL https://github.com/iskandr/fancyimpute.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey, 2022. URL https://arxiv.org/abs/2007.08199.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and

noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21592–21601, 2024.

Indro Spinelli, Simone Scardapane, and Aurelio Uncini. Missing data imputation with adversarially-trained graph convolutional networks. *Neural Networks*, 129:249–260, 2020.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 (1):112–118, 2012.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 06 2001a. ISSN 1367-4803.

Olga G. Troyanskaya, Michael N. Cantor, Gavin Sherlock, Patrick O. Brown, Trevor J. Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17 6:520–5, 2001b.

Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cai Rong Zhao. Learning with noisy labels via self-supervised adversarial noisy masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16186–16195, June 2023.

Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242, 2007.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Eszter Vértes and Maneesh Sahani. Flexible and accurate inference and learning for deep generative models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ramon Vinas, Xu Zheng, and Jer Hayes. A graph-based imputation method for sparse medical records. *arXiv preprint arXiv:2111.09084*, 2021.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Senzhang Wang, Jiyue Li, Hao Miao, Junbo Zhang, Junxing Zhu, and Jianxin Wang. Generative-free urban flow imputation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2028–2037, 2022.

Yufeng Wang, Dan Li, Cong Xu, and Min Yang. Missingness augmentation: A general approach for improving generative imputation models. *arXiv preprint arXiv:2108.02566*, 2021.

Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, June 2022.

Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. Noisywikihow: A benchmark for learning with real-world noisy labels in natural language processing, 2023. URL https://arxiv.org/abs/2305.10709.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

Seongwook Yoon and Sanghoon Sull. Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8456–8464, 2020.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33:19075–19087, 2020.

Muhan Zhang and Yixin Chen. Inductive matrix completion based on graph neural networks. *arXiv preprint arXiv:1904.12058*, 2019.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V. Bonilla. Transformed distribution matching for missing value imputation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42159–42186. PMLR, 23–29 Jul 2023.

Jiajun Zhong, Ning Gui, and Weiwei Ye. Data imputation with iterative graph reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11399–11407, Jun. 2023.

Rui Zhou, Junyan Liu, Sandeep Kumar, and Daniel P Palomar. Student's t var modeling with missing data via stochastic em and gibbs sampling. *IEEE Transactions on Signal Processing*, 68:6198–6211, 2020.

Appendix A. Case Studies

In this section, we conduct case studies comparing our method to the existing Dirichlet regression approach under similar testing conditions, focusing on the ECI of motorcycle accidents, interpersonal violence, car accidents, falls, and firearms.

Motorcycle Accidents On the ECI of motorcycle accidents, determined based on 930,000 injury cases from our dataset, our method improves average RMSE of NI probabilities given the ECI of motorcycle accidents by 26% compared to the existing method. Globally, for the year 2021, the number of head injuries due to motorcycle accidents is estimated to be 932,165 from 6,707,770 motorcycle accidents, comprising 13.89% Institute for Health Metrics and Evaluation (IHME) (2024). For moderate and severe traumatic brain injuries (TBI), an NI in the E-N Matrix, the average RMSE is approximately 4.5% for the existing method and 0.38% for our method, an improvement of 91.7%.

Our method also improves average RMSE for minor TBI by 61.6% compared to the existing method. Given the large number of estimated motorcycle injuries attributable to head injuries in 2021, improving RMSE for moderate and severe TBI by over 90% has a significant impact on making these estimates more accurate. This leads to a more accurate estimation of YLDs and burden of TBI from motorcycle accidents, shaping legislation involving helmets to help reduce the burden of head injuries from motorcycle accidents globally Du et al. (2020).

Interpersonal Violence In the year 2021, the number of incidents of interpersonal violence is estimated to be 29,398,470 globally Institute for Health Metrics and Evaluation (IHME) (2024). Interpersonal violence includes all homicide ECIs relating to physical violence committed against another individual, including homicide by firearm, knife, and others. Based on 1,723,964 injury cases from our dataset, our method reduces average RMSE of NI probabilities given the ECI of interpersonal violence by 70.4%. Following trends over time with regards to interpersonal violence often helps policymakers track successful policies that can help reduce interpersonal violence, rendering these estimations of public health metrics critical for policymaking Mercy et al. (2017).

Car Accidents The ECI of car accidents comprises 1,146,627 injury cases in our dataset and there were an estimated 19,518,750 car accidents in 2021 globally Institute for Health Metrics and Evaluation (IHME) (2024). With our method, the average RMSE of NI probabilities given the ECI of car accidents is improved by 79.4%. Public health metrics focused on car accident related injuries have helped policymakers enact legislation to require car manufacturers to install seat belts Lee et al. (2015); Martin et al. (2018) and airbags Karlson (1992) in their vehicles, reducing the burden of car accidents over the past decades.

Falls The year 2021 was estimated to have over 216 million incidents of falls globally for all ages and sexes Institute for Health Metrics and Evaluation (IHME) (2024). In our dataset, the falls ECI com-

prises 13,485,645 injury cases. Using our method, the average RMSE of NI probabilities given the ECI of falls is reduced by 88.3%. Given the substantial number of incidents of falls estimated for 2021 globally, it is imperative to enact policy to help reduce the burden of falls, which benefits from more accurately estimating public health metrics to determine the largest contributing factors, such as age, sex, and NIs Natora et al. (2022).

Firearms The firearm injury estimation is based on three ECIs, including physical violence by firearm, suicide with a firearm, and an unintentional firearm injury. For the year 2021, there was estimated to be 1,266,596 incidents of physical violence by firearm, 115,867 suicides with a firearm, and 1,785,265 incidents of unintentional firearm injury globally for a total of 3,167,728 firearm incidents Institute for Health Metrics and Evaluation (IHME) (2024). Based on 374,547 injury cases from our dataset, our method improves average RMSE of NI probabilities given the ECI of interpersonal violence by 82.9%.

Firearm related fatal and non-fatal incidents contribute significantly to public health metrics, such as YLLs and DALYs, as they disproportionately affect adolescents and young adults Patel et al. (2022). Regarding the legal purchase of firearms, there is a wide spectrum of policies for firearms globally, such as restrictions on who is allowed to purchase a firearm in a given country or region, in addition to restrictions based on mental health history or criminal background for example Patel et al. (2022); Greenberg et al. (2024). In illegal purchase and ownership of firearms, law enforcement interventions are employed in accordance to policies enacted by different levels of government Patel et al. (2022); Greenberg et al. (2024).

Policies and law enforcement interventions for both the legal and illegal purchase and ownership of firearms can potentially reduce the burden of firearm related incidents Patel et al. (2022). It is critical to provide the most accurate estimates of firearm related public health metrics by country and sub-national level over time to demonstrate the impact and effectiveness of the wide spectrum of firearm policy and intervention.

Appendix B. Loss Evaluation

Self-supervised loss ($\mathcal{L}_D$) was created based on the observation that among certain age groups, NI probabilities are very similar for ECI. This loss is computed using Euclidean distance of NI probabilities for similar age groups (see Section 2.3). To assess the impact of our self-supervised loss, we trained a model with combined loss for 100 epochs after training the model for 50,000 epochs using the MSE loss as part of cold start training discussed in Section 3.2. Our experimental results in Table 2 indicate that larger Θ_D values have a more significant impact on reducing $\mathcal{L}_D$. As indicated in Table 2, the combined loss reduces $\mathcal{L}_D$ from 0.221 to 9.05E-3 after 10 epochs and 7.96E-5 after 100 epochs. However, larger Θ_D values result in an increase of validation and test set RMSE. Further, Θ_D values smaller than 1E-3 fail to reduce the $\mathcal{L}_D$, which is computed as the average Euclidean distance of NI probabilities for similar age groups. Our evaluation results in Table 2 demonstrate that when Θ_D is 5E-3, $\mathcal{L}_D$ drops by 18.5% while only negligibly increasing validations and test set RMSE.

Appendix C. Dataset Applications

In the public health field, our dataset could assist researchers in studying injury patterns, such as the prevalence of various injuries based on age group, sex, and country income status. Beyond the field of injuries and health, learning from noisy labels as well as preventing models from overfitting to noisy samples in the data is a problem of interest in machine learning research in general. Our dataset could be used for a variety of machine learning research, including matrix imputation, deep learning, and regression. These methods are susceptible to overfitting to samples with noisy labels, a common property of real-world data. Datasets that contain noisy labels can be used as benchmarks to develop and evaluate more robust methods against overfitting to samples with noisy labels. However, there is a limited number of datasets containing naturally noisy labels, causing researchers to often add synthetic noise to an otherwise unnoisy dataset Wu et al. (2023). Synthetic noisy labels are only an approximation of naturally noisy labels and can result in machine learning methods performing well on a dataset with synthetic noisy labels and poorly on a real-world dataset with naturally noisy labels Wu et al. (2023). Thus, our dataset, which contains naturally noisy labels, can help make machine learning methods more robust to overfitting to samples with noisy labels. The level of noise in the dataset is also an important factor and datasets with naturally noisy labels that contain over 15% noisy

Table 2: Self-supervised loss $\mathcal{L}_D$ for scale factors Θ_D with validation and test set RMSE

Θ_D	Cold Start			Epoch 10			Epoch 100		
	$\mathcal{L}_D$	Valid	Test	$\mathcal{L}_D$	Valid	Test	$\mathcal{L}_D$	Valid	Test
1.00E-1	2.21E-1	4.24E-2	5.01E-2	9.05E-3	7.57E-2	8.30E-2	7.96E-5	7.61E-2	8.33E-2
7.50E-2	2.23E-1	4.27E-2	5.02E-2	9.18E-3	7.58E-2	8.30E-2	5.21E-6	7.61E-2	8.34E-2
5.00E-2	2.20E-1	4.26E-2	5.05E-2	9.46E-3	7.56E-2	8.28E-2	5.66E-5	7.61E-2	8.33E-2
2.50E-2	2.21E-1	4.27E-2	5.10E-2	3.75E-2	6.82E-2	7.56E-2	4.77E-2	6.10E-2	6.98E-2
2.00E-2	2.21E-1	4.25E-2	5.05E-2	5.84E-2	6.51E-2	7.31E-2	7.57E-2	5.52E-02	6.36E-2
1.50E-2	2.23E-1	4.27E-2	5.02E-2	1.04E-1	5.28E-2	6.06E-2	1.12E-1	4.93E-2	5.79E-2
1.00E-2	2.21E-1	4.25E-2	5.04E-2	1.57E-1	4.58E-2	5.55E-2	1.45E-1	4.57E-2	5.36E-2
7.50E-3	2.20E-1	4.27E-2	5.02E-2	1.76E-1	4.41E-2	5.36E-2	1.58E-1	4.71E-2	5.64E-2
5.00E-3	2.21E-1	4.26E-2	5.05E-2	1.92E-1	4.48E-2	5.49E-2	1.80E-1	4.32E-2	5.16E-2
2.50E-3	2.21E-1	4.25E-2	5.04E-2	2.07E-1	4.39E-2	5.33E-2	2.02E-1	4.28E-2	4.98E-2
1.00E-3	2.22E-1	4.29E-2	5.02E-2	2.18E-1	4.23E-2	5.05E-2	2.15E-1	4.26E-2	5.00E-2
7.50E-4	2.21E-1	4.26E-2	5.03E-2	2.22E-1	4.22E-2	4.95E-2	2.19E-1	4.31E-2	4.98E-2
5.00E-4	2.23E-1	4.26E-2	5.01E-2	2.22E-1	4.31E-2	5.23E-2	2.20E-1	4.26E-2	5.03E-2
2.50E-4	2.22E-1	4.38E-2	5.35E-2	2.22E-1	4.31E-2	5.31E-2	2.22E-1	4.29E-2	5.05E-2
1.00E-4	2.22E-1	4.26E-2	5.04E-2	2.26E-1	4.31E-2	5.24E-2	2.25E-1	4.27E-2	4.99E-2

labels are more uncommon Song et al. (2022). As mentioned in Section 2.2.1, 16% of all samples in our dataset have no injury cases assigned to them, which is incorrect because there cannot be ECIs with all NI probabilities set to zero, thus this is the minimum noise level of our dataset.

Appendix D. Probability Generator

We rely on the probability generator model to learn from our 7 datasets, 6 of which are generated from the original dataset using data augmentation techniques to avoid unevenly learning from the datasets. Further, we use the model to generate probabilities for a multinomial distribution to sample and select a dataset to learn from in a given epoch. The dataset selection probability generator model is an autoencoder used in past works for generative tasks. In Figure 2, we demonstrate the details of the autoencoder architecture of the probability generator model. The generator model uses equal probabilities as inputs and outputs predicted probabilities using the objective function described in Section 2.3.4.

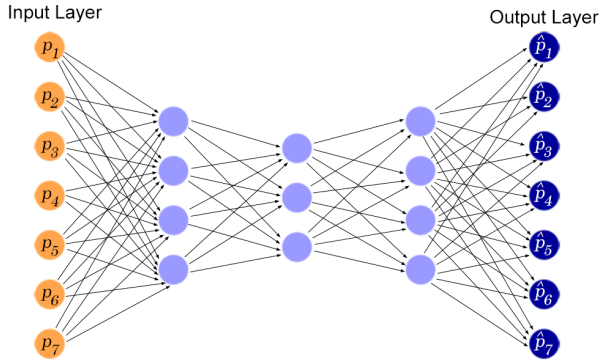

Figure 2: The probability generator model is an autoencoder with ReLU activation function. The model uses equal probabilities as inputs and outputs predicted probabilities using the objective function described in Section 2.3.4.

Appendix E. ECI and NI

This section includes the ICD-9 and ICD-10 codes used to create the NI groupings in Table 3 and ECI groupings in Table 4 for our dataset.

Table 3: Nature-of-Injuries code, description, and ICD codes Haagsma et al. (2016).

Code	**Description**	**ICD-9**	**ICD-10**
N1	Amputation of lower limbs, bilateral	896.2, 896.3, 897.7	T05.3 , T05.5
N2	Amputation of upper limbs, bilateral	887.6, 887.7, 888.1, 888.2, 888.9	S68.4
N3	Amputation of fingers (excluding thumb)	886.0, 886.1	S68.1, S68.6
N4	Amputation of lower limb, unilateral	896.0, 896.1, 897.0, 897.1, 897.2, 897.3, 897.4, 897.5	S78.0, S78.1, S78.9, S88.9, S98.0, S98.3, S98.9
N5	Amputation of upper limb, unilateral	887.0, 887.1, 887.2, 887.3, 887.4, 887.5	S48.9, S58.1
N6	Amputation of thumb	885.0, 885.1	S68.0
N7	Amputation of toe/toes	895.0, 895.1	S98.1
N8	Burns, <20% total burned surface area without lower airway burns	941.0, 941.1, 941.2, 941.3, 941.4, 941.5, 942.0, 942.1, 942.2, 942.3, 943.0, 943.1, 943.2, 943.3, 943.4, 943.5, 944.0, 944.1, 944.2, 944.3, 944.4, 944.5, 945.0, 945.1, 945.2, 945.3, 945.4, 945.5, 947.3, 947.4, 947.8, 947.9, 948.0, 948.1, 949.0, 949.1, 949.2, 949.3, 949.4, 949.5	T20.0, T20.1, T20.2, T20.4, T20.6, T20.7, T21.0, T21.1, T21.2, T21.4, T21.7, T23.1, T24.5, T25.0, T25.1, T25.2, T25.3, T25.4, T25.5, T25.6, T25.7, T28.3, T28.4
N9	Burns, ≥20% total burned surface area or ≥ 10% burned surface area if head/neck or hands/wrist involved w/o lower airway burns	906.5, 906.6, 906.7, 906.8, 906.9, 942.4, 942.5, 946.0, 946.1, 946.2, 946.3, 946.4, 946.5, 948.2, 948.3, 948.4, 948.5, 948.6, 948.7, 948.8, 948.9	T29.6, T31.4, T31.6, T31.8, T31.9, T32.2, T32.4, T32.9
N10	Lower airway burns	947.0, 947.1, 947.2	T27.3
N11	Dislocation of hip	835.0, 835.1	S73.0
N12	Dislocation of knee	836.0, 836.1, 836.2, 836.3, 836.4, 836.5, 836.6	S83.0
N13	Dislocation of shoulder	831.0, 831.1	S43.0, S43.1, S43.2, S43.3

Code	Description	ICD-9	ICD-10
N14	Muscle and tendon injuries, including sprains and strains lesser dislocations	830.0, 830.1, 832.0, 832.1, 832.2, 833.0, 833.1, 834.0, 834.1, 837.0, 837.1, 838.0, 838.1, 839.0, 839.1, 839.2, 839.3, 839.4, 839.5, 839.6, 839.7, 839.8, 839.9, 840.0, 840.1, 840.2, 840.3, 840.4, 840.5, 840.6, 840.7, 840.8, 840.9, 841.0, 841.1, 841.2, 841.3, 841.8, 841.9, 842.0, 842.1, 843.0, 843.1, 843.8, 843.9, 844.0, 844.1, 844.2, 844.3, 844.8, 844.9, 845.0, 845.1, 846.0, 846.1, 846.2, 846.3, 846.8, 846.9, 847.0, 847.1, 847.2, 847.3, 847.4, 847.9, 848.0, 848.1, 848.2, 848.3, 848.4, 848.5, 848.8, 848.9, 905.6, 905.7, 905.8	S03.0, S03.1, S03.2, S03.3, S03.8, S03.9, S13.0, S13.1, S13.2, S13.3, S13.4, S13.5, S13.6, S16.1, S16.2, S16.8, S16.9, S23.0, S23.1, S23.2, S23.3, S23.4, S23.5, S33.3, S43.4, S43.9, S46.1, S46.2, S46.8, S46.9, S53.0, S53.1, S53.4, S66.2, S66.3, S73.1, S76.0, S76.1, S76.2, S76.3, S76.8, S76.9, S86.0, S86.1, S86.2, S86.3, S86.8, S93.0, S93.1, S93.3, S93.4, S93.5, S93.6, S96.0, S96.1, S96.2, S96.9, S99.9
N15	Fracture of clavicle, scapula, or humerus	810.0, 810.1, 811.0, 811.1, 812.0, 812.1, 812.2, 812.3, 812.4, 812.5	S49.0, S49.1
N16	Fracture of face bones	802.0, 802.1, 802.2, 802.3, 802.4, 802.5, 802.6, 802.7, 802.8, 802.9	S02.3, S02.4, S02.5, S02.6, S02.7
N17	Fracture of foot bones except ankle	825.0, 825.1, 825.2, 825.3, 826.0, 826.1, 826.6	S92.3, S92.4, S92.5, S92.7, S92.9
N18	Fracture of hand (wrist and other distal part of hand)	814.0, 814.1, 815.0, 815.1, 816.0, 816.1	S62.8
N19	Fracture of hip	820.0, 820.1, 820.2, 820.3, 820.8, 820.9, 905.3	S72.0, S72.1, S72.2
N20	Fracture of patella, tibia or fibula, or ankle	822.0, 822.1, 823.0, 823.1, 823.2, 823.3, 823.4, 823.8, 823.9, 824.0, 824.1, 824.2, 824.3, 824.4, 824.5, 824.6, 824.7, 824.8, 824.9, 905.4	S82.0, S82.1, S82.2, S82.3, S82.4, S82.5, S82.6, S82.7, S82.8, S82.9, S89.0, S89.1, S89.2, S89.3
N21	Fracture of pelvis	808.0, 808.1, 808.2, 808.3, 808.4, 808.5, 808.8, 808.9	S32.5
N22	Fracture of radius and/or ulna	813.0, 813.1, 813.2, 813.3, 813.4, 813.5, 813.8, 813.9, 905.2	S52.3, S52.4, S52.5, S52.6, S52.7, S59.0, S59.1, S59.2
N23	Fracture of skull	800.0, 800.1, 800.2, 800.3, 800.4, 800.5, 800.6, 800.7, 800.8, 800.9, 801.0, 801.1, 801.2, 801.3, 801.4, 801.5, 801.6, 801.7, 801.8, 801.9, 803.0, 803.1, 803.2, 803.3, 803.4, 803.5, 803.6, 803.7, 803.8, 803.9, 804.0, 804.1, 804.2, 804.3, 804.4, 804.5, 804.6, 804.7, 804.8, 804.9, 905.0	S02.8, S02.9

Code	Description	ICD-9	ICD-10
N24	Fracture of sternum and/or fracture of one or more ribs	807.0, 807.1, 807.2, 807.3, 807.4, 807.5, 807.6	S22.2, S22.3, S22.4, S22.8, S22.9
N25	Fracture of vertebral column	310.2, 805.0, 805.1, 805.2, 805.3, 805.4, 805.5, 805.6, 805.7, 805.8, 805.9, 905.1	S12.0, S12.5, S12.6, S22.0, S22.1
N26	Fracture of femur, other than femoral neck	821.0, 821.1, 821.2, 821.3	S79.1, T93.1
N27	Minor TBI	850.0, 850.1, 850.2, 850.3, 850.4, 850.5, 850.9	G44.3, S06.0
N28	Moderate/Severe TBI	851.0, 851.1, 851.2, 851.3, 851.4, 851.5, 851.6, 851.7, 851.8, 851.9,852.0, 852.1, 852.2, 852.3, 852.4, 852.5, 853.0, 853.1, 854.0, 854.1. 907.0	S06.1, S06.2, S06.3, S06.4, S06.5, S06.6, S06.7, S06.8, S06.9, T90.2
N30	Foreign body in ear	931	T16.0-T16.9
N31	Foreign body in respiratory system	933.0, 933.1, 934.0, 934.1, 934.8, 934.9	T17.2, T17.3, T17.4, T17.8, T17.9
N32	Foreign body in GI and urogenital system	935.0, 935.1, 935.2, 938.9, 939.0, 939.1, 939.2, 939.3, 939.9	T18.1
N33	Spinal cord lesion at neck level	806.0, 806.1, 952.0	S14.1, T91.3
N34	Spinal cord lesion below neck level	806.2, 806.3, 806.4, 806.5, 806.6, 806.7, 806.8, 806.9, 952.1, 952.2, 952.3, 952.4, 952.8, 952.9	S24.0, S24.1, S34.1
N35	Drowning and nonfatal submersion	994.1	T75.1
N36	Asphyxiation	994.7	T71.1, T71.2
N37	Crush injury	906.4, 925.1, 925.2, 926.0, 926.1, 926.8, 926.9, 927.0, 927.1, 927.2, 927.3, 927.8, 927.9, 928.0, 928.1, 928.2, 928.3, 928.8, 928.9, 929.9	S07.0, S07.1, S07.8, S17.0, S17.8, S17.9, S38.0, S67.0, S67.1, S67.3, S67.9, S77.1, S77.2, S87.8, S97.1, S97.8
N38	Nerve injury	907.1, 907.3, 907.4, 907.5, 907.8, 907.9, 950.0, 950.1, 950.2, 950.3, 950.9, 951.0, 951.1, 951.2, 951.3, 951.4, 951.5, 951.6, 951.7, 951.8, 951.9, 953.0, 953.1, 953.2, 953.3, 953.4, 953.5, 953.8, 953.9, 954.0, 954.1, 954.8, 954.9, 955.0, 955.1, 955.2, 955.3, 955.4, 955.5, 955.6, 955.7, 955.8, 955.9, 956.0, 956.1, 956.2, 956.3, 956.4, 956.5, 956.8, 956.9, 957.0, 957.1, 957.8, 957.9	S04.0, S04.1, S04.2, S04.3, S04.4, S04.5, S04.6, S04.7, S04.8, S04.9, S14.2, S14.3, S14.4, S14.5, S14.6, S14.8, S34.8, S34.9, S44.5, S54.0, S54.1, S54.2, S54.3, S64.4, S64.8, S64.9, S74.0, S74.1, S74.2, S74.9, S94.0, S94.1, T13.3, T90.3

Code	Description	ICD-9	ICD-10
N39	Injury to eyes	366.2, 870.0, 870.1, 870.2, 870.3, 870.4, 870.8, 870.9, 871.0, 871.1, 871.2, 871.3, 871.4, 871.5, 871.6, 871.7, 871.9, 918.0, 918.1, 918.2, 918.9, 921.0, 921.1, 921.2, 921.3, 921.9, 930.0, 930.1, 930.2, 930.8, 930.9, 940.0, 940.1, 940.2, 940.3, 940.4, 940.5, 940.9	S01.1, S05.0, S05.1, S05.2, S05.3, S05.4, S05.5, S05.6, S05.7, S05.8, S05.9, T15.0, T15.1, T15.8, T26.4, T26.5, T26.6, T26.8, T90.4
N40	Open wound(s)	872.0, 872.1, 872.6, 872.7, 872.8, 872.9, 873.0, 873.1, 873.2, 873.3, 873.4, 873.5, 873.6, 873.7, 873.8, 873.9, 874.2, 874.3, 874.4, 874.5, 874.8, 874.9, 875.0, 875.1, 876.0, 876.1, 877.0, 877.1, 878.0, 878.1, 878.2, 878.3, 878.4, 878.5, 878.6, 878.7, 878.8, 878.9, 879.0, 879.1, 879.2, 879.3, 879.4, 879.5, 879.6, 879.7, 879.8, 879.9, 880.0, 880.1, 880.2, 881.0, 881.1, 881.2, 882.0, 882.1, 882.2, 883.0, 883.1, 883.2, 884.0, 884.1, 884.2, 890.0, 890.1, 890.2, 891.0, 891.1, 891.2, 892.0, 892.1, 892.2, 893.0, 893.1, 893.2, 894.0, 894.1, 894.2, 900.0, 900.1, 900.8, 900.9, 903.0, 903.1, 903.2, 903.3, 903.4, 903.5, 903.8, 903.9, 904.0, 904.1, 904.2, 904.3, 904.4, 904.5, 904.6, 904.7, 904.8, 904.9, 906.0, 906.1, 906.2	S 01.0, S01.2, S01.3, S01.4, S01.5, S01.7, S01.8, S01.9, S08.0, S08.1, S08.8, S09.0, S09.1, S09.2, S09.3, S10.7, S11.1, S11.8, S11.9, S15.0, S15.1, S15.2, S15.3, S15.7, S15.8, S15.9, S21.0, S21.1, S21.2, S21.3, S21.4, S21.7, S21.8, S21.9, S31.8, S41.0, S41.1, S45.1, S45.3, S51.0, S51.8, S55.0, S55.1, S55.8, S55.9, S65.0, S65.3, S65.4, S65.5, S65.7, S65.8, S65.9, S71.0, S71.1, S71.7, S75.0, S75.1, S75.2, S75.8, S75.9, S81.0, S81.7, S81.8, S81.9, S85.1, S85.2, S85.3, S85.4, S85.5, S85.8, S85.9, S95.0, S95.2, S95.8, T90.1, T93.0

Code	Description	ICD-9	ICD-10
N41	Poisoning requiring urgent care	960.0, 960.1, 960.2, 960.3, 960.4, 960.5, 960.6, 960.7, 960.8, 960.9, 961.0, 961.1, 961.2, 961.3, 961.4, 961.5, 961.6, 961.7, 961.8, 961.9, 962.0, 962.1, 962.2, 962.3, 962.4, 962.5, 962.6, 962.7, 962.8, 962.9, 963.0, 963.1, 963.2, 963.3, 963.4, 963.5, 963.8, 963.9, 964.0, 964.1, 964.2, 964.3, 964.4, 964.5, 964.6, 964.7, 964.8, 964.9, 965.0, 965.1, 965.4, 965.5, 965.6, 965.7, 965.8, 965.9, 966.0, 966.1, 966.2, 966.3, 966.4, 967.0, 967.1, 967.2, 967.3, 967.4, 967.5, 967.6, 967.8, 967.9, 968.0, 968.1, 968.2, 968.3, 968.4, 968.5, 968.6, 968.7, 968.9, 969.0, 969.1, 969.2, 969.3, 969.4, 969.5, 969.6, 969.7, 969.8, 969.9, 970.0, 970.1, 970.8, 970.9, 971.0, 971.1, 971.2, 971.3, 971.9, 972.0, 972.1, 972.2, 972.3, 972.4, 972.5, 972.6, 972.7, 972.8, 972.9, 973.0, 973.1, 973.2, 973.3, 973.4, 973.5, 973.6, 973.8, 973.9, 974.0, 974.1, 974.2, 974.3, 974.4, 974.5, 974.6, 974.7, 975.0, 975.1, 975.2, 975.3, 975.4, 975.5, 975.6, 975.7, 975.8, 976.0, 976.1, 976.2, 976.3, 976.4, 976.5, 976.6, 976.7, 976.8, 976.9, 977.0, 977.1, 977.2, 977.3, 977.4, 977.8, 977.9, 978.0, 978.1, 978.2, 978.3, 978.4, 978.5, 978.6, 978.8, 978.9, 979.0, 979.1, 979.2, 979.3, 979.4, 979.5, 979.6, 979.7, 979.9, 980.0, 980.1, 980.2, 980.3, 980.8, 980.9, 981.2, 981.3, 981.5, 981.6, 981.7, 981.9, 982.0, 982.1, 982.2, 982.3, 982.4, 982.8, 983.0, 983.1, 983.2, 983.5, 983.7, 983.9, 984.0, 984.1, 984.3, 984.8, 984.9, 985.0, 985.1, 985.2, 985.3, 985.4, 985.5, 985.6, 985.8, 985.9, 987.0, 987.1, 987.2, 987.3, 987.4, 987.5, 987.6, 987.7, 987.8, 987.9, 988.0, 988.1, 988.2, 988.6, 988.8, 988.9, 989.0, 989.1, 989.2, 989.3, 989.4, 989.5, 989.6, 989.7, 989.8, 989.9	T36.9, T38.8, T38.9, T39.0, T39.3, T39.8, T39.9, T40.3, T40.4, T40.5, T40.6, T40.9, T41.2, T41.4, T42.7, T43.0, T43.2, T43.4, T43.5, T43.6, T43.9, T44.6, T44.9, T45.5, T45.6, T45.7, T45.8, T45.9, T46.0, T46.1, T46.2, T46.3, T46.4, T46.5, T46.6, T46.7, T46.8, T46.9, T47.0, T47.1, T47.2, T47.3, T47.4, T47.5, T47.6, T47.7, T47.8, T47.9, T48.0, T48.1, T48.2, T48.3, T48.4, T48.5, T48.6, T48.7, T48.9, T49.0, T49.1, T49.2, T49.3, T49.4, T49.5, T49.6, T49.7, T49.8, T49.9, T50.0, T50.3, T50.4, T50.8, T50.9, T51.0, T51.1, T51.2, T52.4, T53.5, T54.9, T56.4, T56.8, T57.9, T58.1, T58.2, T58.8, T58.9, T59.0, T59.1, T59.2, T59.3, T59.4, T59.5, T59.6, T59.9, T60.9, T61.1, T61.7, T62.9, T63.8, T65.2, T65.8, T65.9

Code	Description	ICD-9	ICD-10
N42	Severe chest Injury	860.0, 860.1, 860.2, 860.3, 860.4, 860.5, 861.0, 861.1, 861.2, 861.3, 862.0, 862.1, 862.2, 862.3, 862.8, 862.9, 874.0, 874.1, 901.0, 901.1, 901.2, 901.3, 901.4, 901.8, 901.9, 908.0	S11.0, S11.2, S25.0, S25.1, S25.2, S25.3, S25.4, S25.5, S25.7, S25.8, S25.9, S26.0, S26.1, S27.3, S27.4, S27.8, S28.2, T91.4
N43	Internal hemorrhage in abdomen and pelvis	863.0, 863.1, 863.2, 863.3, 863.4, 863.5, 863.8, 863.9, 864.0, 864.1, 865.0, 865.1, 866.0, 866.1, 867.0, 867.1, 867.2, 867.3, 867.4, 867.5, 867.6, 867.7, 867.8, 867.9, 868.0, 868.1, 868.3, 869.0, 869.1, 902.0, 902.1, 902.2, 902.3, 902.4, 902.5, 902.8, 902.9, 908.1, 908.2, 908.3	S35.1, S35.2, S35.3, S35.4, S35.5, S35.9, S36.0, S36.1, S36.2, S36.3, S36.4, S36.5, S36.6, S36.8, S37.0, S37.2, S37.3, S37.4, S37.5, S37.8, S37.9, T79.6
N44	Contusion in any part of the body	906.3, 922.0, 922.1, 922.2, 922.3, 922.4, 922.8, 922.9, 923.0, 923.1, 923.2, 923.3, 923.8, 923.9, 924.0, 924.1, 924.2, 924.3, 924.4, 924.5, 924.8, 924.9	S20.0, S30.2, S40.2, S50.0, S60.2, S60.8, S70.0, S80.0, S80.1, S80.2, S80.7, S90.0, S90.2
N45	Effect of different environmental factors	991.0, 991.1, 991.2, 991.3, 991.4, 991.5, 991.6, 991.8, 991.9, 992.0, 992.1, 992.2, 992.3, 992.4, 992.5, 992.6, 992.7, 992.8, 992.9, 993.0, 993.1, 993.2, 993.3, 993.4, 993.8, 993.9, 994.0, 994.2, 994.3, 994.4, 994.5, 994.6, 994.8, 994.9	T33.5, T33.8, T34.4, T34.5, T34.6, T34.7, T67.3, T69.0, T69.8, T70.8, T75.2
N46	Complications following therapeutic procedures	995.4, 996.0, 996.1, 996.2, 996.3, 996.4, 996.5, 996.6, 996.7, 996.8, 996.9, 998.0, 998.1, 998.2, 998.3, 998.4, 998.5, 998.6, 998.7, 998.8, 998.9, 999.0, 999.1, 999.2, 999.3, 999.6, 999.7, 999.8, 999.9	T80.3, T80.6, T80.8, T80.9, T81.1, T81.3, T81.5, T81.6, T81.7, T81.8, T82.0, T82.1, T82.2, T82.3, T82.4, T82.5, T82.8, T83.0, T83.1, T83.2, T83.4, T83.7, T83.8, T84.4, T84.8, T85.0, T85.1, T85.2, T85.3, T85.4, T85.5, T85.6, T85.8, T86.1, T86.3, T86.8, T86.9, T87.4, T88.1, T88.2, T88.6, T88.7, T88.8, T88.9

Code	Description	ICD-9	ICD-10
N47	Superficial injury of any part of the body	910.0, 910.1, 910.2, 910.3, 910.4, 910.5, 910.6, 910.7, 910.8, 910.9, 911.0, 911.1, 911.2, 911.3, 911.4, 911.5, 911.6, 911.7, 911.8, 911.9, 912.0, 912.1, 912.2, 912.3, 912.4, 912.5, 912.6, 912.7, 912.8, 912.9, 913.0, 913.1, 913.2, 913.3, 913.4, 913.5, 913.6, 913.7, 913.8, 913.9, 914.0, 914.1, 914.2, 914.3, 914.4, 914.5, 914.6, 914.7, 914.8, 914.9, 915.0, 915.1, 915.2, 915.3, 915.4, 915.5, 915.6, 915.7, 915.8, 915.9, 916.0, 916.1, 916.2, 916.3, 916.4, 916.5, 916.6, 916.7, 916.8, 916.9, 917.0, 917.1, 917.2, 917.3, 917.4, 917.5, 917.6, 917.7, 917.8, 917.9, 919.0, 919.1, 919.2, 919.3, 919.4, 919.5, 919.6, 919.7, 919.8, 919.9	S00.0, S00.1, S00.2, S00.3, S00.4, S00.5, S00.8, S00.9, S10.0, S10.1, S10.8, S10.9, S20.1, S20.3, S20.9, S30.8, S40.2, S40.7, S40.8, S40.9, S50.3, S50.7, S50.8, S70.2, S70.3, S80.8, S80.9, S90.4, S90.5, S90.8, S90.9, T00.8, T00.9, T90.0
N48	Multiple fractures, dislocations, crashes, wounds, pains, and strains	817.0, 817.1, 818.0, 818.1, 819.0, 819.1, 827.0, 827.1, 828.0, 828.1, 929.0	T02.7, T04.7, T06.3
N51	Reserved Codes	-	-
N52	Reserved Codes	-	-
N53	Reserved Codes	-	-
N54	Reserved Codes	-	-
N99	Reserved Codes	-	-

Table 4: External cause of injury code, description, and ICD codes

Code	Description	ICD-9	ICD-10
inj-animal-nonven	Non-venomous animal contact	E906	W52.0-W62.9, W64-W64.9
inj-animal-venom	Venomous animal contact	E905	X20-X29.9
inj-disaster	Exposure to forces of nature	E907-E909	X33-X38.9
inj-drowning	Drowning	E910	W65-W70.9, W73-W74.9
inj-electrocution	Electrocution	E925	W85-W87.9
inj-falls	Falls	E880-E886, E888	W00-W19.9
inj-fires	Fire, heat, and hot substances	E890-E899, E924	X00-X06.9, X08-X19.9
inj-foreign-aspiration	Pulmonary aspiration and foreign body in airway	E911-E913	W75-W75.9, W78-W80.9, W83-W84.9
inj-foreign-eye	Foreign body in eyes	360.5-360.69, 374.86, 376.6, E914-E914.09	H02.81-H02.819, H44.6-H44.799
inj-foreign-other	Foreign body in other body part	E914-E915	W44-W45, W45.3-W45.9
inj-homicide-gun	Physical violence by firearm	E965	X93-X95.9
inj-homicide-knife	Physical violence by sharp object	E966	X99-X99.9
inj-homicide-sexual	Physical sexual violence	E960.0-E960.9	Y05.0-Y05.9
inj-homicide-other	Physical violence by other means	E961-E964, E967-E969	X85-X92.9, X96-X98.9, Y00-Y04.9, Y06-Y08.9, Y87.1
inj-mech-gun	Unintentional firearm injuries	E922	W32-W34.9
inj-mech-other	Other exposure to mechanical forces	E916-E921	W20-W31.9, W35-W38.9, W40-W43.9, W45.0-W45.2, W46-W46.2, W49-W52
inj-non-disaster	Environmental heat and cold exposure	E900-E902, E926 L55-L55.9, L56.3, L56.8-L56.9, L58-L58.9, W88-W94.9,	W97.9, W99-W99.9, X30-X32.9, X39-X39.9

Code	Description	ICD-9	ICD-10
inj-othunintent	Unintentional injuries	349.0-349.1, 457.0, E856-E857, E861-E865, E867-E869, E870-E876, E878-E879, E880-E886, E888-E928, E930-E949	L55-L55.9, L56.3, L56.8-L56.9, L58-L58.9, N30.4, W00-W46.2, W49-W62.9, W64-W70.9, W73-W75.9, W77-W81.9, W83-W94.9, W97.9, W99-X06.9, X08-X39.9, X47-X48.9, X50-X54.9, X57-X58.9, Y40-Y84.9, Y88-Y88.3
inj-poisoning-gas	Poisoning by carbon monoxide	E862, E868-E869	X47-X47.9
inj-poisoning-other	Poisoning by other means	E856-E857, E861, E863-E865, E867	X48-X48.9
inj-suicide-firearm	Self-harm by firearm	E955	X72-X74.9
inj-suicide-other	Self-harm by other specified means	E950-E954, E956-E959	X60-X64.9, X66-X71.9, X75-X83.9, Y87.0
inj-trans-other	Other transport injuries	E800-E807, E830-E838, E840-E849	V00-V00.8, V05-V05.9, V81-V81.9, V83-V86.9, V88.2-V88.3, V90-V98.8
inj-trans-road-2wheel	Motorcyclist road injuries	E810.2, E810.3, E811.2, E811.3, E812.2, E812.3, E813.2, E813.3, E814.2, E814.3, E815.2, E815.3, E816.2, E816.3, E817.2, E817.3, E818.2, E818.3, E819.2, E819.3, E820.2, E820.3, E821.2, E821.3, E822.2, E822.3, E823.2, E823.3, E824.2, E824.3, E825.2, E825.3	V20-V29.9
inj-trans-road-4wheel	Motor vehicle road injuries	E810.0, E810.1, E811.0, E811.1, E812.0, E812.1, E813.0, E813.1, E814.0, E814.1, E815.0, E815.1, E816.0, E816.1, E817.0, E817.1, E818.0, E818.1, E819.0, E819.1, E820.0, E820.1, E821.0, E821.1, E822.0, E822.1, E823.0, E823.1, E824.0, E824.1, E825.0, E825.1	V30-V79.9, V87.2-V87.3
inj-trans-road-other	Other road injuries	E810.4, E810.5, E811.4, E811.5, E812.4, E812.5, E813.4, E813.5, E814.4, E814.5, E815.4, E815.5, E816.4, E816.5, E817.4, E817.5, E818.4, E818.5, E819.4, E819.5, E820.4, E820.5, E821.4, E821.5, E822.4, E822.5, E823.4, E823.5, E824.4, E824.5, E825.4, E825.5, E826.3, E826.4, E827.3, E827.4, E828.4, E829.4	V80-V80.9, V82-V82.9

Code	Description	ICD-9	ICD-10
inj-trans-road-pedal	Cyclist road injuries	E800.3, E801.3, E802.3, E803.3, E804.3, E805.3, E806.3, E807.3, E810.6, E811.6, E812.6, E813.6, E814.6, E815.6, E816.6, E817.6, E818.6, E819.6, E820.6, E821.6, E822.6, E823.6, E824.6, E825.6, E826.1	V10-V19.9
inj-trans-road-pedest	Pedestrian road injuries	E810.7, E811.7, E812.7, E813.7, E814.7, E815.7, E816.7, E817.7, E818.7, E819.7, E822.7, E823.7, E824.7, E825.7, E826.0, E827.0, E828.0, E829.0	V01-V04.9, V06-V09.9
inj-war-execution	Police conflict and executions	E970-E978	Y35-Y35.9, Y89.0
inj-war-terrorism	Conflict and terrorism	E979, E990-E999	U00-U03, Y36-Y38.9, Y89.1

Appendix F. Larger Age Groups

In our approach, larger age groupings were used for data augmentation and self-supervised loss calculation. Among the larger age groupings, NI probabilities given ECI are expected to be similar, according to field experts, including public health scientists and clinicians contributing to this study. The larger age groupings, combining the age groups included in the dataset and the age ranges of individuals associated with injury cases in the dataset, are provided in Table 5.

Appendix G. Vetting Groups

The larger groupings of ECIs for vetting are included in Table 6. The description of ECI codes used in Table 6 can be found in Table 4.

Appendix H. Hyperparameters

The hyperparameters we used for our experiments are gathered in Table 7, in an effort to make our contributions reproducible.

Table 5: Larger age grouping for data augmentation

Age Group ID	**Age Groups**	**Age Range**
1	0, 1	[0, 5)
15	5, 10, 15	[5, 20)
35	20, 25, 30, 35	[20, 40)
70	40, 45, 50, 55, 60, 65, 70	[40, 75)
95	75, 80, 85, 90, 95+	[75, 95+]

Table 6: External causes of injuries groups for vetting

Group Names	**External Causes of Injuries Codes**
Poison	inj-animal-venom, inj-poisoning-gas, inj-poisoning-other
Physical Trauma	inj-animal-nonven, inj-disaster, inj-falls, inj-foreign-eye, inj-foreign-other, inj-mech-other, inj-homicide-knife, inj-homicide-other, inj-othunintent, inj-suicide-other, inj-trans-road-2wheel, inj-war-execution, inj-trans-road-4wheel, inj-trans-road-other, inj-trans-road-pedal, inj-trans-road-pedest, inj-trans-other, inj-war-terrorism
Burn	inj-fires, inj-electrocution, inj-non-disaster, inj-war-terrorism, inj-disaster, inj-trans-road-2wheel, inj-trans-road-4wheel, inj-war-execution
Violence	inj-homicide-gun, inj-mech-gun, inj-suicide-firearm, inj-war-execution, inj-war-terrorism, inj-homicide-knife, inj-homicide-other
Firearm	inj-homicide-gun, inj-mech-gun, inj-suicide-firearm, inj-war-execution, inj-war-terrorism
Strangulation	inj-foreign-aspiration, inj-drowning
Other	inj-non-disaster

Table 7: Final hyperparameters

Method	**Hyperparameters**
Random	Implementation: Fancyimpute Rubinsteyn and Feldman, fill-method: "random", min-value: 0.0, max-value: 1.0
Mean	Implementation: Fancyimpute Rubinsteyn and Feldman, fill-method: "mean", min-value: 0.0, max-value: 1.0
Dirichlet RG (current)	NA
KNN	Implementation: Fancyimpute Rubinsteyn and Feldman, k: 5
MICE	Implementation: Fancyimpute Rubinsteyn and Feldman, IterativeImputer
EM	Implementation: Impyute Law (2017), eps: 0.6
SoftImpute	Implementation: Fancyimpute Rubinsteyn and Feldman, convergence-threshold: 0.00001, max-iters: 1000, max-rank: 57
IterativeSVD	Implementation: Fancyimpute Rubinsteyn and Feldman, rank: 56, convergence-threshold: 0.0000001, max-iters: 1000
Matrix Factorization	Implementation: Fancyimpute Rubinsteyn and Feldman, rank: 40, learning-rate: 0.0002, max-iters: 20, shrinkage-value: 0.0, min-value: 0.0, max-value: 1
GAIN	Implementation: GAIN for Pytorch Brahma (2024), Mini batch size: 2048, Hint rate: 1.0, Alpha: 10
Round Robin MLP	Implementation: OT Imputer Muzellec et al. (2020), Random Seed: 1234, batchsize: 256, lr:5e-4
MIRACLE (Mean)	Implementation: MIRACLE Kyono et al. (2021), Seed: Predicted E-N matrix using the Mean method, n-hidden: 64, reg-m: 0.1, lr: 0.00005, window: 10, max-steps: 5
MIRACLE (GRAPE)	Implementation: MIRACLE Kyono et al. (2021), Seed: Predicted E-N matrix using the GRAPE method, n-hidden: 64, reg-m: 0.1, lr: 0.00005, window: 10, max-steps: 5
OT Imputer	Implementation: OT Imputer Muzellec et al. (2020), Random Seed: 1234, batchsize: 256, lr:5e-4, niter: 50000
TDM	Implementation: TDM Zhao et al. (2023), Random Seed: 1234, niter: 10000, batchsize: 1024, lr: 5e-4, noise: 0.09, network-depth: 4, network-width: 3
IGRM	Implementation: IGRM Zhong et al. (2023), weight-decay: 0.0000001, lr: 0.002, known: 0.6, valid: 0.125, epochs: 50000, opt-scheduler: step, opt-decay-step: 500, opt-decay-rate: 0.94, node-dim: 128, edge-dim: 128, impute-hiddens: 128, dropout: 0.1, aggr: mean
GRAPE	Implementation: GRAPE You et al. (2020), weight-decay: 0.0000001, lr: 0.002, known: 0.5, valid 0.125, epochs: 50000: opt-scheduler: step, opt-decay-step: 500, opt-decay-rate: 0.94, node-dim: 256, edge-dim: 256, impute-hiddens: 256, aggr: mean
Self-supervised GRAPE(pre-training)	weight-decay: 0.0000001, lr: 0.00125, known: 0.425, valid: 0.125, epochs: 50100, opt-scheduler: step, opt-decay-step: 800, opt-decay-rate: 0.94, node-dim: 256, edge-dim: 256, impute-hiddens: 256, aggr: mean, dropout: 0.05, dist-loss-cold-start-delay 50000, dist-loss-delta 0.0002, dist-loss-proba: 1.00, dist-loss-iter: 10
Self-supervised GRAPE(fine-tuning)	different parameters from pre-training(lr: 0.00035, epochs: 20100, opt-decay-step: 250)