Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: A Retrospective Observational Study

Maruthi Kumar Mutnuri, Henry Thomas Stelfox, Nils Daniel Forkert, Joon Lee

Submitted to: Journal of Medical Internet Research
on: September 14, 2023

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: A Retrospective Observational Study

Maruthi Kumar Mutnuri[1, 2]; Henry Thomas Stelfox[3, 4] MD, PhD; Nils Daniel Forkert[5, 6] PhD; Joon Lee[1, 7, 8, 9] PhD

[1]Data Intelligence for Health Lab Cumming School of Medicine University of Calgary Calgary CA
[2]Department of Biomedical Engineering Schulich School of Engineering University of Calgary Calgary CA
[3]Department of Critical Care Medicine Cumming School of Medicine University of Calgary Calgary CA
[4]O'Brien Institute for Public Health Cumming School of Medicine University of Calgary Calgary CA
[5]Department of Radiology Cumming School of Medicine University of Calgary Calgary CA
[6]Alberta Children's Hospital Research Institute Cumming School of Medicine University of Calgary Calgary CA
[7]Department of Cardiac Sciences Cumming School of Medicine University of Calgary Calgary CA
[8]Department of Community Health Sciences Cumming School of Medicine University of Calgary Calgary CA
[9]Department of Preventive Medicine School of Medicine Kyung Hee University Seoul KR

Corresponding Author:
Joon Lee PhD
Data Intelligence for Health Lab
Cumming School of Medicine
University of Calgary
CWPH 5E17
3280 Hospital Drive NW
Calgary
CA

Abstract

Background: Accurate patient outcome prediction in the intensive care unit (ICU) can lead to more effective and efficient patient care. Deep learning models are capable of learning from data to accurately predict patient outcomes, but they typically require large amounts of data and computational resources. Transfer learning (TL) can help in scenarios when data and computational resources are scarce by leveraging pre-trained models. While TL has been widely used in medical imaging and natural language processing, it has been rare in electronic health record (EHR) analysis. Furthermore, domain adaptation (DA) has been the most commonly used TL method in general, whereas inductive transfer learning (ITL) has been rare. To the best of our knowledge, DA and ITL have never been studied in depth in the context of EHR-based ICU patient outcome prediction.

Objective: This study investigated DA as well as rarely researched ITL in EHR-based ICU patient outcome prediction under simulated, varying levels of data scarcity.

Methods: Two patient cohorts were used in this study: 1) eCritical, a multicenter ICU data from 55,689 unique admission records from 48,672 unique patients admitted to 15 medical-surgical ICUs in Alberta, Canada, between March 2013 and December 2019; and 2) MIMIC-III, a single-center, publicly available ICU dataset from Boston, USA, acquired between 2001 and 2012. We compared DA and ITL models with baseline models (without TL) of fully connected neural networks, logistic regression, and lasso regression in the prediction of 30-day mortality, acute kidney injury (AKI), ICU length of stay (ICU_LOS), and hospital length of stay (H_LOS). Random subsets of training data, ranging from 1% to 75%, as well as the full dataset were used to compare the performances of DA and ITL with the baseline models at various levels of data scarcity.

Results: Overall, the ITL models outperformed the baseline models in 55 out of 56 comparisons. The DA models outperformed the baseline models in 45 out of 56 comparisons. ITL resulted in better performance than DA in terms of the number of times and the margin with which it outperformed the baseline models. In 11 out of 16 cases (8 out of 8 for ITL and 3 out of 8 for DA), TL models outperformed baseline models when trained using the 1% data subset.

Conclusions: TL-based ICU patient outcome prediction models are useful in data-scarce scenarios. The results of the present study can be used to estimate ICU outcome prediction performance at different levels of data scarcity, with and without TL. The publicly available pre-trained models from this study can serve as building blocks in further research for the development and validation of models in other ICU cohorts and outcomes.

(JMIR Preprints 14/09/2023:52730)
DOI: https://doi.org/10.2196/preprints.52730

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Using Domain Adaptation and Inductive Transfer Learning to Improve Patient Outcome Prediction in the Intensive Care Unit: A Retrospective Observational Study

Abstract

Background: Accurate patient outcome prediction in the intensive care unit (ICU) can potentially lead to more effective and efficient patient care. Deep learning models are capable of learning from data to accurately predict patient outcomes, but they typically require large amounts of data and computational resources. Transfer learning (TL) can help in scenarios where data and computational resources are scarce by leveraging pre-trained models. While TL has been widely used in medical imaging and natural language processing, it has been rare in electronic health record (EHR) analysis. Furthermore, domain adaptation (DA) has been the most common TL method in general, whereas inductive transfer learning (ITL) has been rare. To the best of our knowledge, DA and ITL have never been studied in depth in the context of EHR-based ICU patient outcome prediction.
Objective: This study investigated DA as well as rarely researched ITL in EHR-based ICU patient outcome prediction under simulated, varying levels of data scarcity.
Methods: Two patient cohorts were used in this study: 1) eCritical, a multicenter ICU data from 55,689 unique admission records from 48,672 unique patients admitted to 15 medical-surgical ICUs in Alberta, Canada, between March 2013 and December 2019; and 2) MIMIC-III, a single-center, publicly available ICU dataset from Boston, USA, acquired between 2001 and 2012 containing 61,532 admission records from 46,476 patients. We compared DA and ITL models with baseline models (without TL) of fully connected neural networks, logistic regression, and lasso regression in the prediction of 30-day mortality, acute kidney injury (AKI), ICU length of stay (ICU_LOS), and hospital length of stay (H_LOS). Random subsets of training data, ranging from 1% to 75%, as well as the full dataset were used to compare the performances of DA and ITL with the baseline models at various levels of data scarcity.
Results: Overall, the ITL models outperformed the baseline models in 55 out of 56 comparisons (all p-values < 0.001). The DA models outperformed the baseline models in 45 out of 56 comparisons (all p-values < 0.001). ITL resulted in better performance than DA in terms of the number of times and the margin with which it outperformed the baseline models. In 11 out of 16 cases (8 out of 8 for ITL and 3 out of 8 for DA), TL models outperformed baseline models when trained using the 1% data subset.
Conclusions: TL-based ICU patient outcome prediction models are useful in data-scarce scenarios. The results of the present study can be used to estimate ICU outcome prediction performance at different levels of data scarcity, with and without TL. The publicly available pre-trained models from this study can serve as building blocks in further research for the development and validation of models in other ICU cohorts and outcomes.

Keywords: transfer learning; patient outcome prediction; intensive care; deep learning; electronic health record

Introduction

Electronic Health Records (EHRs) are databases that hospitals and healthcare providers use to record an individual's health history. There has been significant progress in using deep learning models for predicting patient outcomes using EHR data [1]. However, using deep learning models is not feasible in some settings such as rural hospital ICUs, which have low patient volumes and limited computational capacity due to budget restrictions.

Transfer learning (TL) can be useful in these challenging scenarios. TL research using EHR data has been uncommon compared to medical image analysis and natural language processing. The basic idea of TL is to utilize the knowledge and representations learned while training the model on a source prediction task, to improve prediction performance on a different, but potentially closely related target prediction task [2]. In practice, this is achieved by pre-training a model with data for the source prediction task and saving the trained weights. These weights capture the intrinsic knowledge of the data. This pre-trained model is typically loaded without the final layers (usually the fully connected layers immediately preceding the output layer and the output layer itself), which are then replaced with new layers. Finally, this pre-trained model is retrained to fine-tune it to the target prediction task (Figure 1).

In this work, we considered two commonly used types of TL methods. First, inductive transfer learning (ITL) aims to improve performance on the target task after learning a different but related source task, usually from the same domain [3]. For example, in Tokuoka *et al.* [4] a model trained for brain tissue annotation label (source task) was adapted to the task of brain tumor segmentation (target task) using Magnetic Resonance Imaging images. Second, transductive transfer learning, or also referred to as domain adaptation (DA), makes use of different domains but the same prediction task [3]. For example, in Titoriya et al. [5] AlexNet [6] model which was pre-trained on ImageNet data (source dataset) for object classification (source task) was re-trained on the Breakhis dataset (target dataset) to classify medical images into malignant or benign (target task) to predict breast cancer.

TL can be useful in a data-scarce scenario where the target dataset does not have a sufficient volume to train a deep learning model but there is a sufficiently large source dataset to train a pre-trained model or a relevant pre-trained model is available. For example, DA can be useful in a scenario where a rural intensive care unit (ICU) does not have sufficient data to train a model to predict a patient outcome, but an urban teaching hospital has a large dataset to train a model for the same patient outcome. ITL can be useful when predicting new or rare patient outcomes at an ICU if that ICU has a sufficiently large dataset for training a model to predict a different patient outcome.

TL research in medical image analysis is relatively comprehensive, as pre-trained convolutional neural networks (CNN) such as AlexNet [6], ResNet [7], VGGNet [8], and GoogleNet [9] are publicly available and have been used widely for prediction problems such as image classification [10], image segmentation [11], object identification [12], disease categorization [13], and severity grading [14]. Most of these examples employed DA rather than ITL.

Another field with significant previous TL research is natural language processing. Some of the established pre-trained language models are Word2Vec [15], GloVe [16], BERT [17], and fastText [18]. Some of the use cases of pre-trained natural language processing models include text mining [19], word classification [20], and sentiment classification [21].

A recent trend has been to adapt the progress made in TL research in the natural language processing domain to EHR data analysis. Inspired by the pre-trained BERT model [17], Li *et al.* [22], developed BEHRT using EHR data to improve future visit diagnosis prediction. This study used only four features, which included age, segment, position, and diagnosis. Longitudinal data were used, and a transformer-based model was trained. This study had inclusion criteria of patients with at least 5 visits and a diagnosis available in their EHR. One of the shortcomings of this study was not utilizing all the available features in the EHR data such as demographics, lab results, vitals, prescriptions, etc., Another concern was the requirement of five previous visits with diagnosis, causing the model to be unable to predict the diagnosis for patients with fewer previous visits.

Liu *et al.* [23] used TL to improve the prediction of acute kidney injury (AKI) using EHR data acquired at the University of Kansas Medical Center. Logistic regression (LR) was used as the global (baseline) model and the global TL model. Global TL and baseline models are the same LR models except the baseline model was trained on the original dataset and the TL model was trained on the modified dataset. To create the modified dataset, data of each feature were multiplied by the corresponding feature coefficient, which was obtained from the global baseline model. The personalized model was then trained on a subset of the training sample with the highest similarity (nearest neighbors) with the selected test sample. For each test sample, a subset of the training dataset with the highest k-NN score was selected and a personalized model and personalized TL model were trained. The personalized TL model followed the same approach as the global TL model except that the training sample was selected from the modified dataset using the k-NN score (similarity) with the selected test sample. Although the source and target prediction tasks and domains were the same, this can be considered as DA since TL models use a modified dataset. Here, deep learning models were not used, which are often assumed to be better at learning the general representations of the data. Since both these TL models (global and personalized) were not like the traditional pre-trained models with saved weights, the transfer of knowledge happened by modifying the dataset. Thus, the transferred knowledge is stored in a modified dataset, not in the TL models. These TL models are strongly tied to this specific dataset and their generalizability has not been well established by using external validation data.

In Shickel *et al.* [24], TL was used to improve hospital discharge prediction using a conventional ICU cohort acquired at the University of Florida Health as the source cohort and the Intelligent ICU cohort also acquired at the University of Florida Health as the target cohort. It utilized a feed-forward neural network for TL. Here, the source cohort had 48,400 patient records whereas the target cohort had 51 patient records. This study used only 9 features and DA, even though not stated specifically in the manuscript. This semantic progression was observed in other studies as well; DA has become so prevalent that the terms TL and DA are being used interchangeably, with the former most commonly referring to the latter [25]-[24].

To the best of our knowledge, TL research has been rare in EHR-based ICU patient outcome prediction, particularly in terms of ITL, leading to a limited understanding of the effectiveness of TL in data-scarce ICU settings. Furthermore, there is a lack of publicly available EHR-based pre-trained models for ICU patient outcome prediction. These are important gaps because EHR-based tabular data are one of the most widely-used forms of data in predictive modeling studies in health, and the state-of-the-art TL methods from general computer vision and natural language processing are often not readily applicable to EHR data. Hence, this retrospective study aimed to compare the performances of DA, ITL, and baseline (without TL) models in predicting the following four ICU patient outcomes at varying levels of data scarcity: 30-day post-ICU admission mortality, acute kidney injury (AKI), hospital length of stay (H_LOS), and ICU length of stay (ICU_LOS).

Methods

Data Sources

EHR data from two patient cohorts were used in this retrospective study. The first cohort was eCritical, which has 55,689 unique admission records from 48,672 unique patients admitted to 15

ICUs in Alberta, Canada, between March 2013 and December 2019. The second cohort was the MIMIC-III database (Medical Information Mart for Intensive Care) [26] version 1.4, which includes 61,532 unique admission records from 46,476 unique patients admitted to the ICUs at the Beth Israel Deaconess Medical Center in Boston, MA, USA, between 2001 and 2012.

Research Ethics

Since MIMIC-III is a publicly available database, the need to obtain research ethics approval to use it in this study was waived. However, eCritical contains patient identifying information and approval from the Conjoint Health Research Ethics Board, University of Calgary, was obtained (REB17-0389). Informed consent was waived due to the large number of patients involved in the study. All research was performed in accordance with relevant guidelines and regulations set by the University of Calgary and Alberta Health Services, the custodian of the eCritical data, as well as the Declaration of Helsinki.

Patient Cohorts

Two different sets of inclusion and exclusion criteria were applied. The first set was used to establish the base cohorts for the entire study, whereas the second set was specific to each patient outcome and applied to the base cohorts.

For the base cohorts, the following inclusion criteria were applied to both eCritical and MIMIC-III (Figure 2): 1) only the first ICU admission of each patient; 2) ICU length of stay greater than 24 hours; 3) only adult patients with an age 18 years and over; and 4) samples with data available of at least 80% of the features (missing values are imputed as discussed in data preprocessing section). As a result, the base cohorts for eCritical and MIMIC-III consisted of 39,317 and 31,446 patient records, respectively.

Further inclusion and exclusion criteria were applied to each patient outcome, resulting in different datasets for each outcome, as shown in Figure 2. 30-day mortality did not require further inclusion criteria and was modeled using the base cohorts. For AKI, the following criteria were applied: 1) only patient records with sufficient data to determine the presence or absence of AKI, with one serum creatinine lab measurement within the first 24 hours of admission to be able to establish a baseline and another measurement after 24 hours of ICU admission; and 2) no AKI onset at or within 24 hours of admission. For ICU_LOS, the following inclusion criteria were applied: 1) the presence of ICU admission and discharge date-times; and 2) to exclude outliers only the bottom 98th percentile values of ICU_LOS are included. For H_LOS, the following inclusion criteria were applied: 1) the presence of hospital admission and discharge date-times; and 2) to exclude outliers only the bottom 98th percentile values of H_LOS were included.

In the end, the eCritical and MIMIC-III cohorts had 39,317 and 31,446 samples respectively for 30-day mortality. Similarly, AKI had 32,076 and 26,741 samples, H_LOS had 37,675 and 30,816 samples, and ICU_LOS had 38,529 and 30,816 samples, respectively. These cohorts were randomly split into 80% training, 10% validation, and 10% test data. We compared the prediction performance of TL with those of baseline models at different levels of training data scarcity with random subsets of 1%, 5%, 10%, 25%, 50%, 75%, and 100% of the training data.

Patient Outcomes

The primary patient outcomes for this study were 30-day post-ICU admission mortality and ICU_LOS. A patient was defined as deceased if he/she died after being admitted to the ICU and within 30 days of ICU admission. ICU_LOS was defined as the time between ICU admission and discharge.

Furthermore, AKI after 24 hours of ICU admission and H_LOS were predicted as secondary patient outcomes. AKI was identified using the creatinine criteria of KDIGO [27]. H_LOS was defined as the time between hospital admission and discharge.

30-day mortality and AKI were predicted as classification problems, whereas H_LOS and ICU_LOS were predicted as regression problems.

Feature Set

Machine learning (ML) models were trained with the following predictor variables from the first 24 hours in the ICU that were common in both eCritical and MIMIC-III: demographics, vitals, laboratory test results, Glasgow Coma Scale (GCS), prescriptions, dialysis, and mechanical ventilation. Features with more than 30% missing data were excluded from the study. Table 1 shows a complete list of the predictor variables. Four statistical features (5th percentile, 95th percentile, Interquartile Range (IQR), and median) were extracted from the longitudinal variables such as vitals, lab results, and GCS. The maximum and minimum values of labs and vitals carry crucial information regarding the health condition of the patient, but to minimize the influence of outliers, the 5 percentile and 95 percentiles were used instead. Other predictor variables, such as prescriptions, dialysis, and mechanical ventilation, were transformed into binary features to indicate the presence or absence. In the end, a total of 104 features were included in this study.

Table 1. Common feature set between eCritical and MIMIC-III. GCS: Glasgow Coma Scale; BP: blood pressure; RBC: red blood cell; WBC: white blood cell

Feature	Category	Unit of Measurement
Age	demographics	years
Weight	demographics	Kg
Sex	demographics	Binary (M/F)
Eye Opening	GCS	
Verbal Response	GCS	
GCS	GCS	
Motor Response	GCS	
Urine Volumes	Urine Volumes	mL
Heart Rate	Vitals	bpm
BP Systolic	Vitals	mmHg
BP Diastolic	Vitals	mmHg
SpO2	Vitals	%
Respiratory Rate	Vitals	breaths/min
Urea Blood	Labs	mmol/L
CO2 Content Blood	Labs	mmol/L
Creatinine Blood	Labs	umol/L
Glucose Blood	Labs	mmol/L
Potassium Blood	Labs	mmol/L
Sodium Blood	Labs	mmol/L
PCO2 Arterial	Labs	mmHg
FiO2	Labs	%
PH Arterial	Labs	
PO2 Arterial	Labs	mmHg

Hemoglobin	Labs	g/L
Hematocrit	Labs	%
RBC	Labs	E+12 units/L
WBC	Labs	E+9 units/L
Dialysis	Dialysis	Binary (1/0)
Mechanical Ventilation	Mechanical Ventilation	Binary (1/0)
Norepinephrine	Prescriptions	Binary (1/0)
Phenylephrine	Prescriptions	Binary (1/0)
Vasopressin	Prescriptions	Binary (1/0)
Dobutamine	Prescriptions	Binary (1/0)
Dopamine	Prescriptions	Binary (1/0)
Epinephrine	Prescriptions	Binary (1/0)

Data Preprocessing

Differences in units of measurement between eCritical and MIMIC-III were handled by converting all features in MIMIC-III to the eCritical units of measurement.

The train set was used for training the models, whereas the validation set was used for tuning hyperparameters. The test set was used for model performance evaluation. Numerical features (e.g., vitals, labs) were scaled (unit variance and zero mean) and categorical features (e.g., sex, prescriptions, mechanical ventilation) were transformed using one-hot encoding.

Missing data were present to varying degrees in both cohorts. Features with more than 30% missing data were excluded from the study. Patient records with missing data for more than 20% of the features were dropped. The remaining patient records with missing values were imputed using the IterativeImputer from the scikit-learn Python package, which is similar to the multiple imputation by chained equations. Imputation was performed after splitting the data to avoid data leakage. Training, validation, and test data were imputed separately.

Both categorical patient outcomes, 30-day mortality, and AKI, had varying degrees of class imbalance in both cohorts. 30-day mortality had the highest class imbalance; the event rates were 16.79% and 12.24% in eCritical and MIMIC-III, respectively. The class imbalance was mitigated using SMOTE [28], by over-sampling the minority class to 50% of the majority class and then under-sampling the majority class to 100% of the minority class.

Baseline Models

Since logistic [29] and lasso [30] regression are widely used in medical research for classification and regression, respectively, they were used as baseline models. Also, deep learning models (FCNN) with random initialization of weights were used as baseline models.

Hyperparameter tuning for the LR and lasso models was done using grid search with 3-fold cross-validation. The searched hyperparameter space for LR included: solvers of newton-cg, liblinear, lbfgs, sag, and saga; penalties of L1, L2, and none; and C values ranging from 0.01 to 10 with a step of 0.01. The searched hyperparameter space for lasso included: alpha values ranging from 0.01 to 1 with a step of 0.01. FCNN baseline models used the same architecture and hyperparameters as the corresponding TL models so that we were comparing models that were trained the same way except how the weights were initialized.

Eight FCNN models were created for the four patient outcomes and two cohorts. Four LR models were created for 30-day mortality and AKI trained on eCritical and MIMIC-III. Similarly, four lasso regression models were created for H_LOS and ICU_LOS trained on eCritical and MIMIC-III.

Transfer Learning Models

In DA, the source and target domains were eCritical and MIMIC-III, respectively. The source and target tasks were the same, and each DA model predicted one of the four patient outcomes. Each model was pre-trained on the source domain data before being fine-tuned and evaluated on the target domain data. As a result, four pre-trained DA models were created, one for each of the four patient outcomes.

In ITL, both the source and target domains were eCritical and the source task was 30-day mortality prediction whereas the target task was the prediction of one of the four patient outcomes. The ITL model where both the source and target tasks were 30-day mortality served as a benchmark for the other ITL models. In the end, four pre-trained ITL models were created for each of the four patient outcomes.

The pre-trained TL models were fully connected neural networks trained on the training dataset of the source domain. Hyperparameters were tuned using the validation dataset. The searched hyperparameter space included: dropout rates of 0.5, 0.4, and 0.3; batch sizes of 32, 64, and 128, numbers of neurons per hidden layer of 100, 128, 256, and 200; learning rates of 0.001 and 0.0001; activation functions of ReLU, tanh, selu, elu, LeakyReLU, and PReLU; kernel initializers of HeUniform and HeNormal; and kernel regularizers of L2 (l2=1e-3) and L1 (l1=0.001). Also, different architectures were explored. The first one had three hidden layers with layer, layer/2, and layer/4 number of neurons. The second architecture was seven hidden layers with layer, layer*2, layer*2, layer, layer, layer/2, and layer/4 neurons. Here, the layer had 100, 200, 128, and 256 neurons. Finally, these models were tested using the hold-out test set from the source domain to identify the best performing model concerning balanced accuracy (to account for class imbalance) for classification tasks and mean absolute error (MAE) for regression tasks. Then, these best performing models were used as the pre-trained models.

For fine-tuning, the pre-trained model was loaded and the last hidden layer was replaced with a new hidden layer with randomly initialized weights. Then, all pre-trained model layers were frozen (preventing those layers from learning) except for the newly added hidden layer and the model was trained to allow the new hidden layer to adjust its weights. Then, all layers were unfrozen (allowing weights to update) and the model was trained for the final time.

Prediction Performance Comparisons at Varying Levels of Data Scarcity

To investigate prediction performance at varying levels of data scarcity, random subsets of 1%, 5%, 10%, 25%, 50%, and 75% were created from the full training dataset (100%). To avoid selection bias [31], each subset was obtained 10 different times using 10 different random states. Models were trained on these 10 data subsets and then performance metrics from all models were then aggregated for each subset. As there were six subsets (1%, 5%, 10%, 25%, 50%, and 75%) and the full training dataset of 100%, 61 (6x10+1) models were trained for each outcome and each model. For example, for AKI, 61 LR, 61 FCNN, 61 ITL, and 61 DA models were trained.

To obtain the median and 95% confidence intervals of the performance metrics, 1000 bootstrap samples of the test set were obtained for the full dataset (100%), and for the random subsets, 100 bootstrap samples for each of the 10 random states (1000 in total) were created and then tested using these bootstrapped test sets.

All classification models were assessed using balanced accuracy as the primary metric (to account for class imbalance) and the following four secondary metrics on the hold-out test set: area under the receiver operating characteristic curve (AUC), accuracy, precision (a.k.a., positive predictive value), and recall (a.k.a., sensitivity). All regression models were evaluated using MAD and mean squared error (MSE).

Finally, Wilcoxon rank sum tests were performed to compare the performance of TL models to the baseline models. Since there were repeated comparisons involved, a Bonferroni correction was applied. Because the classification tasks had 35 comparisons (7 data subsets and 5 metrics), statistical significance was indicated by p < 0.0014 (0.05 / 35). The regression tasks had 14 comparisons (7 data subsets and 2 metrics), leading to statistical significance set at p < 0.0035 (0.05 / 14).

Results

Patient Cohorts

Based on the inclusion and exclusion criteria, the final eCritical and MIMIC-III cohorts were different for each patient outcome (Figure 2). The 30-day mortality cohort had 39,317 and 31,446 samples in eCritical and MIMIC-III databases, respectively, whereas the AKI cohort had 32,076 and 26,741 samples, respectively. The H_LOS cohort had 37,675 and 30,816 samples, and the ICU_LOS cohort had 38,529 and 30,816 samples, respectively. In the eCritical cohort, there were 6,713 (17.07%) 30-day mortalities, whereas, in the MIMIC-III database, there were 3,900 (12.40 %). The eCritical cohort had 4,524 (14.11 %) AKI cases whereas MIMIC-III had 5,789 (21.64 %). The eCritical cohort had a median H_LOS of 11.48 days with an interquartile range (IQR) of (5.59, 23.29), whereas the MIMIC-III cohort had a median (IQR) of 7.39 (4.67, 12.32). Similarly, the median (IQR) ICU_LOS in the eCritical and MIMIC-III cohorts were 3.97 (2.2, 7.67) and 2.47 days (1.59, 4.58), respectively. The descriptive statistics for the two cohorts are shown in Table 2.

Table 2: Descriptive statistics of the two patient cohorts.

Descriptor	eCritical	MIMIC-III
Male, n (%)	22,957 (58.39)	17,900 (56.92)
Age (years), median (IQR)	60 (46, 70)	66 (53, 78)
Admission weight (Kg), median (IQR)	80 (67.4, 96.6)	79.3 (66.5, 94)
30-day mortality, n (%)	6,713 / 39,317 (17.07 %)	3,900/ 31,446 (12.40 %)
AKI, n (%)	4,524/ 32,076 (14.10 %)	5,789/ 26,741 (21.64 %)
H_LOS (days), median (IQR)	11.48 (5.59, 23.29)	7.39 (4.67, 12.32)
ICU_LOS (days), median (IQR)	3.97 (2.2, 7.67)	2.47 (1.59, 4.58)
Blood creatinine (µmol/L), median (IQR)	90.5 (65.5, 152.05)	79.56 (61.88, 123.76)
Glasgow Coma Scale, median (IQR)	11 (7.2, 14.25)	12(8, 15)
Blood glucose (mmol/L), median (IQR)	7.49 (6.09, 9.45)	7.1 (5.83, 8.99)
Blood potassium (mmol/L), median (IQR)	3.95 (3.6, 4.4)	4.07 (3.7, 4.5)
Blood sodium (mmol/L), median	138.2 (135.15, 141)	138.5 (136, 141)

(IQR)		
Arterial PH, median (IQR)	7.38 (7.32, 7.43)	7.38 (7.33, 7.43)
WBC (K/uL), median (IQR)	12.03 (8.55, 16.65)	11.2 (8.1, 14.9)
RBC (m/uL), median (IQR)	3.7 (3.15, 4.24)	3.54 (3.15, 3.99)
Systolic blood pressure (mmHg), median (IQR)	118.5 (100.25, 139.8)	117 (100.8, 135.75)
Diastolic blood pressure (mmHg), median (IQR)	61.5 (52, 73.8)	59.5 (49.5, 71)
SpO2 (%), median (IQR)	97 (94, 99)	97.7 (95, 99.7)
Hemoglobin (g/L), median (IQR)	112.45 (95.75, 129.2)	107 (94.5, 121)
Hematocrit (%), median (IQR)	34 (29, 39)	31.5 (27.95, 35.5)
Mechanical ventilation, n (%)	34,073 (86.66)	16,429 (52.25)
Dialysis, n (%)	2,173 (5.53)	6,909 (21.97)
Norepinephrine, n (%)	15,797 (40.18)	3,339 (10.62)
Phenylephrine, n (%)	4,384 (11.15)	6,737 (21.42)
Vasopressin, n (%)	5087 (12.94)	712 (2.26)
Dobutamine, n (%)	754 (1.92)	429 (1.36)
Dopamine, n (%)	1051 (2.67)	1512 (4.81)
Epinephrine, n (%)	1320 (3.36)	1221 (3.88)

Pre-trained Models

After hyperparameter tuning for the 30-day mortality source task, the pre-trained model with three hidden layers of 128, 64, and 32 neurons was selected. This model had the highest balanced accuracy of 0.7810. This was the pre-trained model for all four ITL target tasks and the 30-day mortality DA target task. Similarly, after hyperparameter tuning for the AKI, H_LOS, and ICU_LOS source tasks, pre-trained models with three hidden layers of 256, 128, and 64 neurons, seven hidden layers of 256, 512, 512, 256, 256, 128, and 64 neurons, and seven hidden layers of 256, 512, 512, 256, 256, 128, and 64 neurons, with a balanced accuracy of 0.7199, an MAE of 11.8019, and an MAE of 3.0887, were selected, respectively. These were the pre-trained models for the DA target tasks for AKI, H_LOS, and ICU_LOS, respectively.

These pre-trained models are publicly available via GitHub [32].

Domain Adaptation

Multimedia Appendices 1-4 show the complete prediction performances of the DA and baseline models at varying levels of data scarcity represented by the data subsets for 30-day mortality, AKI, ICU_LOS, and H_LOS, respectively. Figures 3-6 pictorially compare the DA and baseline models for each patient outcome.

For 30-day mortality, DA models outperformed both the baseline models LR and FCNN for data subsets 1% to 50%. For datasets, 75% and 100% DA model outperformed the LR model but underperformed the FCNN model. For example, when 1% dataset was used for training, DA model had a median balanced accuracy of 0.6744 (95% CI: 0.5758,0.7083), whereas LR had 0.5821 (0.551,0.6134) and FCNN had 0.6636 (0.6146,0.6971).

For AKI, the DA models outperformed both baseline models for some of the data subsets (75%, 50%, 25%, 10%, and 5%) and underperformed both baseline models for the data subset (1%). The DA model outperformed the LR model and underperformed the FCNN model for the 100% dataset. When the 10% data subset was used, the DA model had a median balanced accuracy of 0.6511 (95% CI: 0.626,0.6763), whereas the LR model had 0.6052 (0.5779,0.6262) and the FCNN model had 0.6439 (0.6177,0.6678).

For ICU_LOS, the DA models outperformed both baseline models for some of the data subsets (25% to 100%). The DA models outperformed the FCNN models but underperformed the lasso models in some cases (5%, and 10%). Also, the results from the 1% data subset were not significant between DA and FCNN (p=0.0468). When the 25% training data subset was used for training, the DA model had a median MAE of 2.2781(95% CI: 2.0427,4.643), whereas the lasso model had 2.4165(2.2967,11.1001) and the FCNN model had 3.8481(3.5641,5.0331).

For H_LOS, the DA models outperformed both baseline models for some of the data subsets (25% to 100%). The DA models outperformed the FCNN models but underperformed the lasso models in some cases (1% and 5%). The results from the 10% data subset were not significantly different between DA and lasso (p=0.0193), but the DA model outperformed the FCNN model. When the 25% data subset was used for training, the DA model had a median MAE of 4.9109 (95% CI: 4.5982,7.389), whereas the lasso model had 5.0491 (4.8903,6.457) and the FCNN model had 9.2677 (9.0162,9.6332).

Inductive Transfer Learning

Multimedia Appendices 5, 6, 3, and 4 show the complete prediction performances of the ITL and baseline models at varying levels of data scarcity represented by the data subsets for 30-day mortality, AKI, ICU_LOS, and H_LOS, respectively. Figures 3-6 pictorially compare the ITL and baseline models for each patient outcome.

As mentioned in the Methods, the 30-day mortality prediction results from ITL presented in Multimedia Appendix 5 serve as a benchmark, since the source and target domains and prediction tasks were the same. The fast convergence of ITL performance at very small data subsets shown in Figure 3 corroborates the limited learning taking place during fine-tuning.

For AKI, the ITL models outperformed both baseline models for all data subsets except 100%. For example, when the 1% dataset was used for training the models, the ITL model had a median balanced accuracy of 0.6434 (95% CI: 0.6006,0.6888), whereas the LR model had 0.5467 (95% CI: 0.5154,0.5732) and the FCNN model had 0.6222 (95% CI: 0.5757,0.6604).

For ICU_LOS, the ITL models outperformed both baseline models for all the data subsets. For example, when the 1% dataset was used, the ITL model had a median MAE of 3.4519 (95% CI: 3.2863,3.8158), whereas lasso had 3.5883 (95% CI: 3.4255,3.7376) and FCNN had 5.626 (95% CI:5.4351,5.8329).

For H_LOS, the ITL models outperformed both baseline models for all data subsets. For example, when the 1% dataset was used, the ITL model had a median MAE of 13.3182 (95% CI: 12.6128,13.9609) whereas lasso had 13.7765 (95% CI: 13.3118,14.2661) and FCNN had 18.5363 (95% CI: 17.8711,19.243).

Discussion

Principal Results

Overall, the ITL models outperformed the baseline models in 55 out of the 56 cases (7 data subsets x 4 outcomes x 2 baseline models). The DA models outperformed the baseline models in 45 times out of 56 cases. While TL is expected to yield better prediction performance than the baseline models, particularly with small data subsets, this assumption has not been confirmed in a comprehensive manner yet in the context of EHR-based ICU patient outcome prediction. In particular, the ITL prediction performances reported in this study are important contributions, given that ITL has seldom been investigated in EHR-based studies in general.

Moreover, the results from this study characterize DA, ITL, and baseline prediction performances as a function of target data volume. ICUs with limited data or computing resources can use our results as a guide to decide which would be better: fine-tune our pre-trained models on their data or train new models from scratch using their small data set.

It is also worth noting that DA did not always outperform the baseline models even at very small data subsets (e.g., ICU_LOS at the 1% subset; see Multimedia Appendix 3). This finding implies that one should not blindly apply TL even when the target data set is small and expect performance improvement.

The ITL models performed better than the DA models in terms of the number of times and the margin with which they outperformed the baseline models. This speaks to the fact that the eCritical and MIMIC-III cohorts are quite different, and the knowledge learned from eCritical exhibited limited utility in predicting the outcomes of the MIMIC-III patients. This is corroborated by the substantial differences in all four outcomes shown in Table 2. There seem to be more similarities between different outcomes within the same cohort than between different cohorts for the same outcome. This finding may be surprising to many researchers since the differences between different patient outcomes in terms of disease progression and risk factors are believed to be substantial, whereas even at different sites the fundamentals of the diseases and ICU care should have many similarities.

Our pre-trained models have been made available publicly which can be used at other ICUs or in future research. In many scenarios, TL was useful even at the 1% data subset, representing only about 200 samples. Hence, ICUs with their own data set containing just 200 samples can potentially benefit from our pre-trained models. Given the paucity of public pre-trained models in EHR-based ICU patient outcome prediction, our pre-trained models are an invaluable contribution to the field.

Clinical Implications

While accurate patient outcome prediction in the ICU has the potential to enable early initiation of preventative care and improve clinical efficiency and care resource management, it may be infeasible for many ICUs to build their own predictive models due to a lack of digital data infrastructure. The pre-trained models from this study address this barrier by serving as predictive models that can be used out-of-the-box (albeit sub-optimal performance) or fine-tuned with a small amount of local data. This study provides a pathway for a wider group of ICUs to consider bringing patient outcome prediction models to the point of care.Comparisons with the Literature

This study has several strengths in comparison with previous EHR-based TL studies. First, our study is one of the few studies that explored ITL using EHR data. Second, our study used two large cohorts: eCritical with 55,689 ICU admissions from 48,672 patients as the source domain, and MIMIC-III with 61,532 ICU admissions from 46,476 patients as the target domain. These are considerably larger than the cohorts used in previous DA studies. For example, Shickel *et al.* [24] used the conventional ICU cohort at the University of Florida Health with 48,400 distinct ICU admissions as the source domain and the Intelligent ICU cohort at the University of Florida Health with only 51 ICU admissions as the target cohort. Third, our study used a large feature set of 104. In comparison, Shickel et al. [24] used only 9 features, while Li et al. [22] used 4 features. Fourth, our pre-trained models are able to predict outcomes for new patients without previous admissions, unlike the natural language processing models developed by Li et al. [22].

Secondary Performance Metric Results

While the TL models outperformed the baseline models in general with respect to the primary performance metrics (balanced accuracy and MAE), the baseline models (particularly LR and lasso) often outperformed the TL models in terms of the secondary metrics. In mortality and AKI prediction, the LR models tended to show higher precisions and lower recalls than the TL models. Given that both mortality and AKI exhibited low event rates leading to substantial class imbalance, the precision and recall results indicate that the LR models were more biased toward the majority class than the TL models. The higher AUCs from the LR models indicate that they achieved higher specificities in general than the TL models, further corroborating the bias toward the majority class. This is why we chose balanced accuracy as our primary metric since it reflects performances on both the majority and minority classes.

In the regression tasks of ICU_LOS and H_LOS prediction, the lasso models often yielded better results than the TL models in terms of MSE. This implies that the TL models often led to larger errors that were amplified by the squaring effect of MSE. Similarly, the large confidence intervals in Figures 5(B) and 6(B) are also likely due to occasional large errors caused by the fact that the model output is not upper-bounded.

Limitations

This study has limitations. First, we could not include all available features due to the constraint of having to use common features in both eCritical and MIMIC-III. Second, this study could not investigate other major ICU patient outcomes such as sepsis, delirium, and acute respiratory distress syndrome, due to data unavailability in either or both eCritical and MIMIC-III. Third, only two baseline models were investigated per prediction task and more advanced ML models (e.g., XGBoost) were not used. However, because the primary objective of our study was to demonstrate the benefits of TL, the most appropriate benchmark models were the FCNNs where everything was equivalent to the TL models except for the use of a pre-trained model. While our focus was not to produce the best prediction performance, our results are comparable to the best MIMIC-based results in the literature as shown by the review study conducted by Syed et al. [33]. Fourth, this study only examined the discrimination of the prediction models and did not investigate calibration. While many health ML studies focus only on discrimination and neglect calibration [34], this remains an important limitation of this study. Lastly, even though some of the features such as vitals and labs were longitudinal, we performed a cross-sectional study via feature aggregation. More advanced recurrent deep learning models (e.g., long short-term memory, gated recurrent unit) that can leverage longitudinal information may have led to different results.

Future Work

First, future work can include an investigation of DA and ITL on other ICU patient outcomes such as sepsis and delirium using data sets that can support such research. Second, application of DA and

ITL to more advanced recurrent deep learning models would be worthwhile. Third, the effectiveness of pre-trained models in the combination of both DA and ITL (i.e., both the domain and prediction task would change from source to target) remains to be studied. For example, a mortality prediction model pre-trained on a source data set can be fine-tuned on a target data set from a different domain to predict AKI. Lastly, the effectiveness of TL across patient subgroups with respect to demographics and socioeconomic status would be worthwhile investigating.

Conclusions

In this retrospective study, we found that TL can lead to improved prediction performance when compared to baseline models trained from scratch only using target data. This performance improvement was observed at a wide range of simulated data scarcity. Also, the performance of ITL was superior to that of DA. This implies fine-tuning a pre-trained model to predict a different patient outcome within the same domain would be a promising approach. We hope that the pre-trained models from this study are useful to other researchers and ICUs.

Acknowledgements

MM was supported by an Alberta Innovates Graduate Student Scholarship for Data-Enabled Innovation from Alberta Innovates and Alberta Advanced Education, as well as a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-02588). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

MM designed the study, prepared the data, conducted all experiments, and wrote the manuscript. JL conceived and designed the study, and directly supervised MM. HS facilitated accessing the eCritical data. All authors contributed to revisions of the manuscript.

The authors would like to thank Dr. Filipe Lucini for assistance with data retrieval, and Alberta Health Services for granting access to the eCritical and administrative health databases. Finally, ARC and MARC clusters used in this project were provided by the University of Calgary.

Data and Code Availability

The eCritical data contain patient identifiable information and cannot be made publicly available. MIMIC is a publicly available database and access can be obtained [35].

The code that conducted the experiments as well as the pre-trained models are available on GitHub [32].

Conflicts of Interest

None declared.

Abbreviations

AKI: acute kidney injury
AUC: area under the receiver operating characteristic curve
BA: balanced accuracy
CI: confidence interval
DA: domain adaptation
EHR: electronic health record

FCNN: fully connected neural network
GCS: Glasgow Coma Scale
H_LOS: hospital length of stay
ICU: intensive care unit
ICU_LOS: intensive care unit length of stay
ITL: inductive transfer learning
IQR: interquartile range
LR: logistic regression
MAE: mean absolute error
ML: machine learning
MSE: mean squared error
TL: transfer learning

Multimedia Appendices

Multimedia Appendix 1: 30-day mortality prediction performances of DA and baseline models across all data subsets.

Multimedia Appendix 2: AKI prediction performances of DA and baseline models across all data subsets.

Multimedia Appendix 3: ICU_LOS prediction performances of ITL, DA, and baseline models across all data subsets.

Multimedia Appendix 4: H_LOS prediction performances of ITL, DA, and baseline models across all data subsets.

Multimedia Appendix 5: 30-day mortality prediction performances of ITL and baseline models across all data subsets.

Multimedia Appendix 6: AKI prediction performances of ITL and baseline models across all data subsets.

References

1. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589–1604. doi: 10.1109/JBHI.2017.2767063

2. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 2013 Aug;35(8):1798–1828. doi: 10.1109/TPAMI.2013.50

3. Niu S, Liu Y, Wang J, Song H. A Decade Survey of Transfer Learning (2010–2020). IEEE Trans Artif Intell 2020 Oct;1(2):151–166. doi: 10.1109/TAI.2021.3054609

4. Tokuoka Y, Suzuki S, Sugawara Y. An Inductive Transfer Learning Approach using Cycle-consistent Adversarial Domain Adaptation with Application to Brain Tumor Segmentation. Proc 2019 6th Int Conf Biomed Bioinforma Eng New York, NY, USA: ACM; 2019. p. 44–48. doi:

10.1145/3375923.3375948

5. Titoriya A, Sachdeva S. Breast Cancer Histopathology Image Classification using AlexNet. 2019 4th Int Conf Inf Syst Comput Netw ISCON IEEE; 2019. p. 708–712. doi: 10.1109/ISCON47742.2019.9036160

6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May;60(6):84–90. doi: 10.1145/3065386

7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016;2016-Decem:770–778. doi: 10.1109/CVPR.2016.90

8. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014 Sep; doi: https://doi.org/10.48550/arXiv.1409.1556

9. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions. 2014 Sep; Available from: http://arxiv.org/abs/1409.4842

10. Ananda Kumar KS, Prasad AY, Metan J. A hybrid deep CNN-Cov-19-Res-Net Transfer learning architype for an enhanced Brain tumor Detection and Classification scheme in medical image processing. Biomed Signal Process Control 2022 Jul;76:103631. doi: 10.1016/j.bspc.2022.103631

11. Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, Guttmann CRG, de Leeuw F-E, Tempany CM, van Ginneken B, Fedorov A, Abolmaesumi P, Platel B, Wells WM. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. 2017. p. 516–524. doi: 10.1007/978-3-319-66179-7_59

12. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016 May;35(5):1285–1298. doi: 10.1109/TMI.2016.2528162

13. Diamant I, Bar Y, Geva O, Wolf L, Zimmerman G, Lieberman S, Konen E, Greenspan H. Chest Radiograph Pathology Categorization via Transfer Learning. Deep Learn Med Image Anal Elsevier; 2017. p. 299–320. doi: 10.1016/B978-0-12-810408-8.00018-3

14. Sugeno A, Ishikawa Y, Ohshima T, Muramatsu R. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. Comput Biol Med 2021 Oct;137:104795. doi: 10.1016/j.compbiomed.2021.104795

15. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 Jan; Available from: http://arxiv.org/abs/1301.3781

16. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proc 2014 Conf Empir Methods Nat Lang Process EMNLP Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. p. 1532–1543. doi: 10.3115/v1/D14-1162

17. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 Oct; Available from:

http://arxiv.org/abs/1810.04805

18. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText.zip: Compressing text classification models. 2016 Dec; Available from: http://arxiv.org/abs/1612.03651

19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, editor. Bioinformatics 2020 Feb;36(4):1234–1240. doi: 10.1093/bioinformatics/btz682

20. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep Learning--based Text Classification. ACM Comput Surv 2022 Apr;54(3):1–40. doi: 10.1145/3439726

21. Gao Z, Feng A, Song X, Wu X. Target-Dependent Sentiment Classification With BERT. IEEE Access 2019;7:154290–154299. doi: 10.1109/ACCESS.2019.2946594

22. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. Sci Rep 2020 Apr;10(1):7155. doi: 10.1038/s41598-020-62922-y

23. Liu K, Zhang X, Chen W, Yu ASL, Kellum JA, Matheny ME, Simpson SQ, Hu Y, Liu M. Development and Validation of a Personalized Model With Transfer Learning for Acute Kidney Injury Risk Estimation Using Electronic Health Records. JAMA Netw Open 2022 Jul;5(7):e2219776. doi: 10.1001/jamanetworkopen.2022.19776

24. Shickel B, Davoudi A, Ozrazgat-Baslanti T, Ruppert M, Bihorac A, Rashidi P. Deep Multi-Modal Transfer Learning for Augmented Patient Acuity Assessment in the Intelligent ICU. Front Digit Health 2021 Feb;3. doi: 10.3389/fdgth.2021.640685

25. Harsono IW, Liawatimena S, Cenggoro TW. Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning. J King Saud Univ - Comput Inf Sci 2022 Mar;34(3):567–577. doi: 10.1016/j.jksuci.2020.03.013

26. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data Nature Publishing Groups; 2016;3(1):1–9. PMID:27219127

27. Kellum, J. A. ; Lameire N. Kidney disease: Improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl 2012 Mar;2(1):1–138. doi: 10.1038/kisup.2012.1

28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–357. doi: 10.1613/jair.953

29. Ge W, Huh J-W, Park YR, Lee J-H, Kim Y-H, Turchin A. An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. AMIA Annu Symp Proc AMIA Symp 2018;2018:460–469. PMID:30815086

30. Hepp T, Schmid M, Gefeller O, Waldmann E, Mayr A. Approaches to Regularized Regression – A Comparison between Gradient Boosting and the Lasso. Methods Inf Med 2016 May;55(05):422–430. doi: 10.3414/ME16-01-0033

31. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection Bias and Information Bias in Clinical

Research. Nephron Clin Pract 2010 Apr;115(2):c94--c99. doi: 10.1159/000312871

32. Mutnuri M. ICU_outcome_prediction_Transfer_Learning. Available from: https://github.com/data-intelligence-for-health-lab/ICU_outcome_prediction_Transfer_Learning [accessed Sep 12, 2023]

33. Syed M, Syed S, Sexton K, Syeda HB, Garza M, Zozus M, Syed F, Begum S, Syed AU, Sanford J, Prior F. Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review. Informatics 2021 Mar;8(1):16. doi: 10.3390/informatics8010016

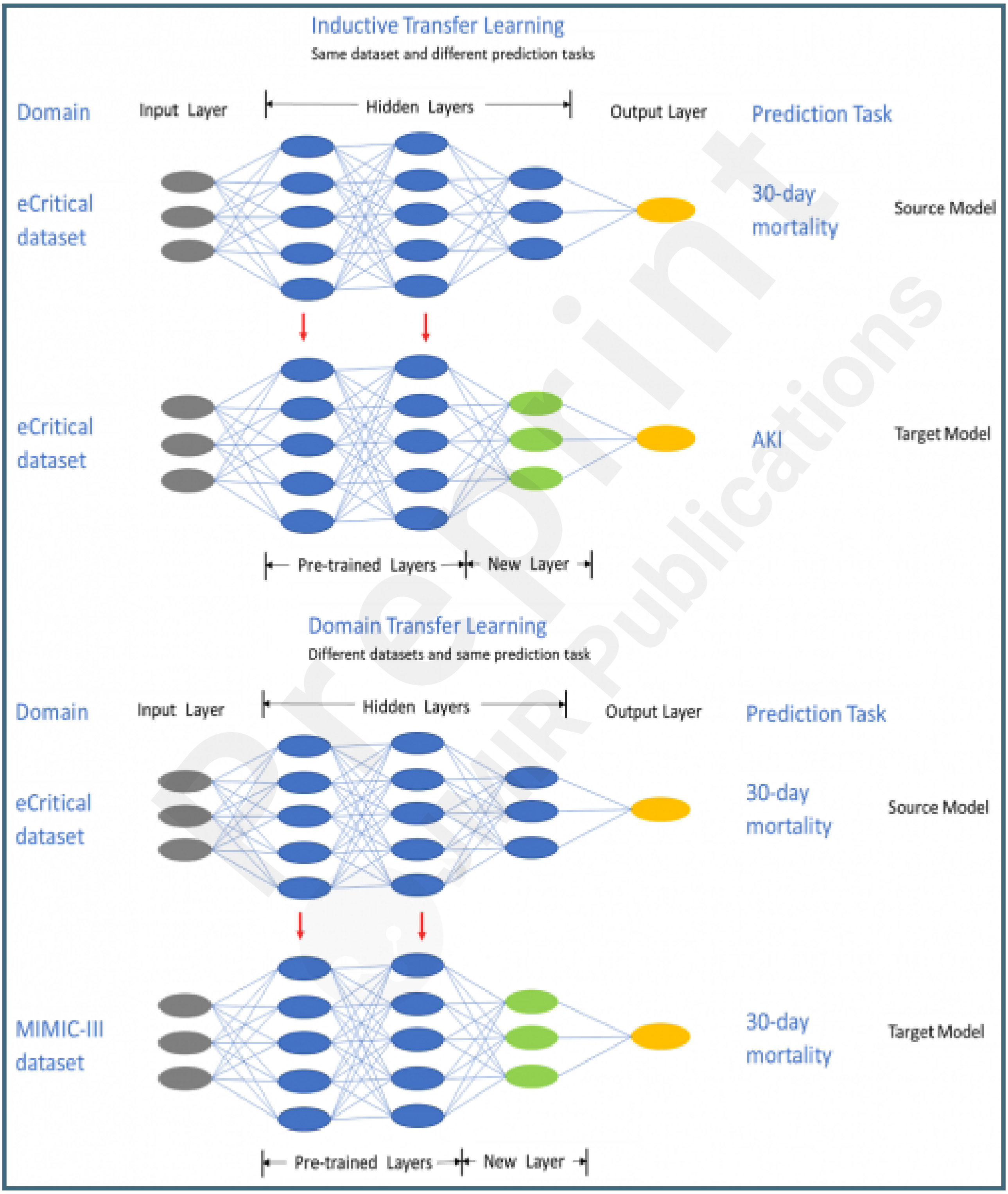
34. Staartjes VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning–based predictive analytics. J Neurosurg Spine 2020 Feb 21;32(6):985–987. doi: 10.3171/2019.12.SPINE191503

35. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database. Available from: https://physionet.org/content/mimiciii/1.4/ [accessed Sep 12, 2023]

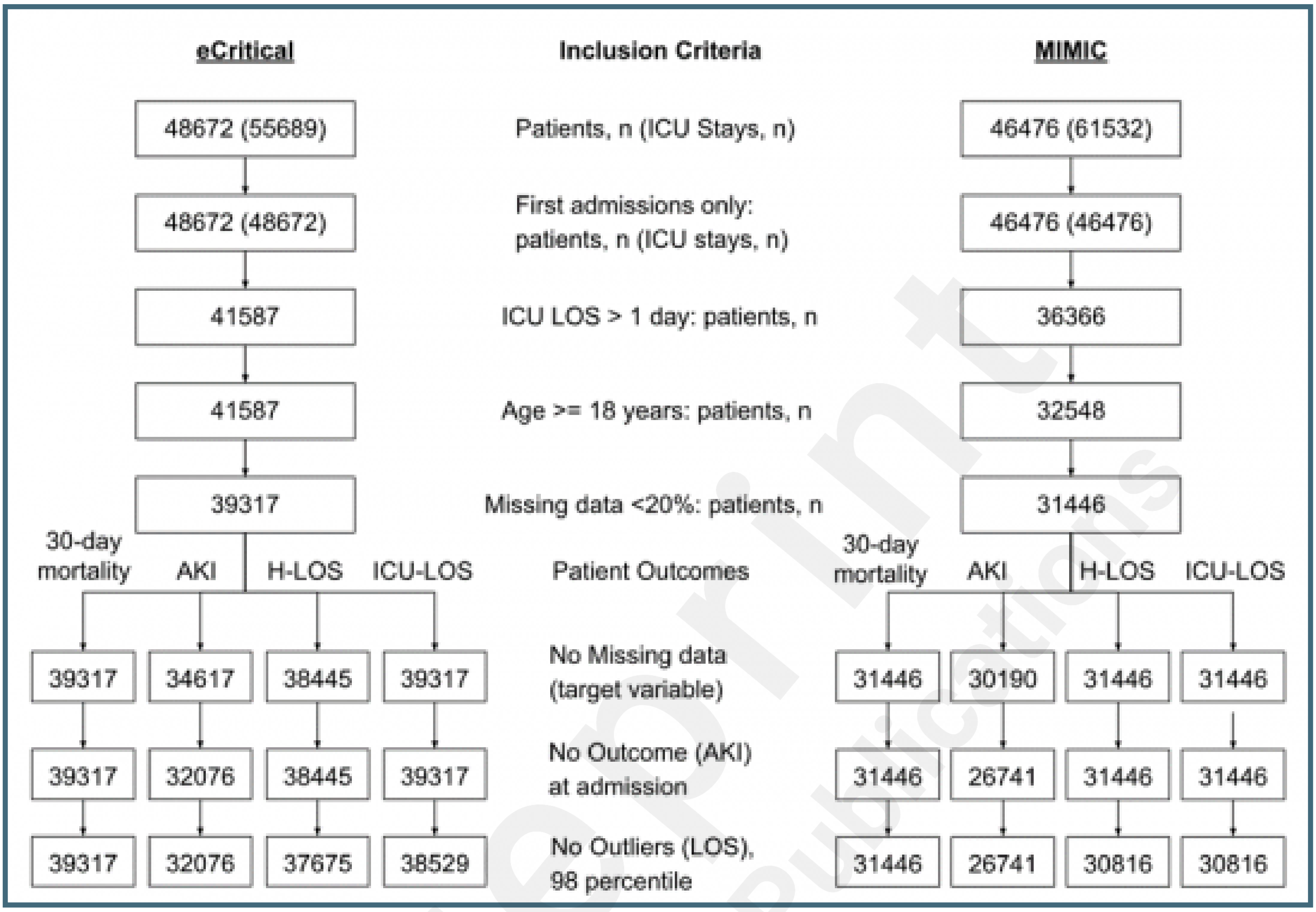
Supplementary Files

Figures

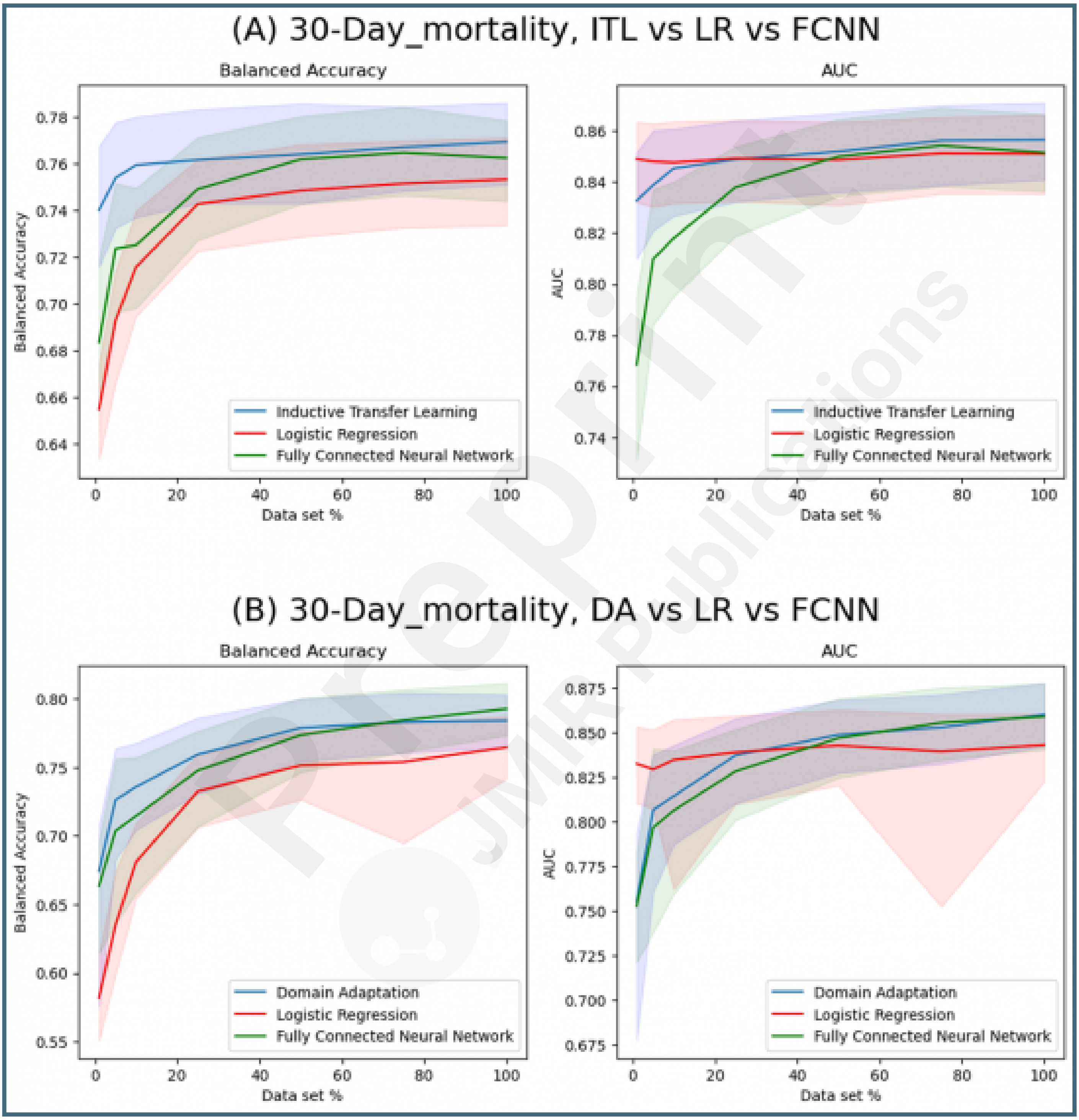
An illustration of how domain adaptation and inductive transfer learning were applied in this study. For inductive transfer learning, the source and target prediction tasks were 30-day mortality and AKI, respectively, while the source and target domains were both eCritical. For domain adaptation, the source and target prediction tasks were both 30-day mortality, while the source and target domains were eCritical and MIMIC-III, respectively.



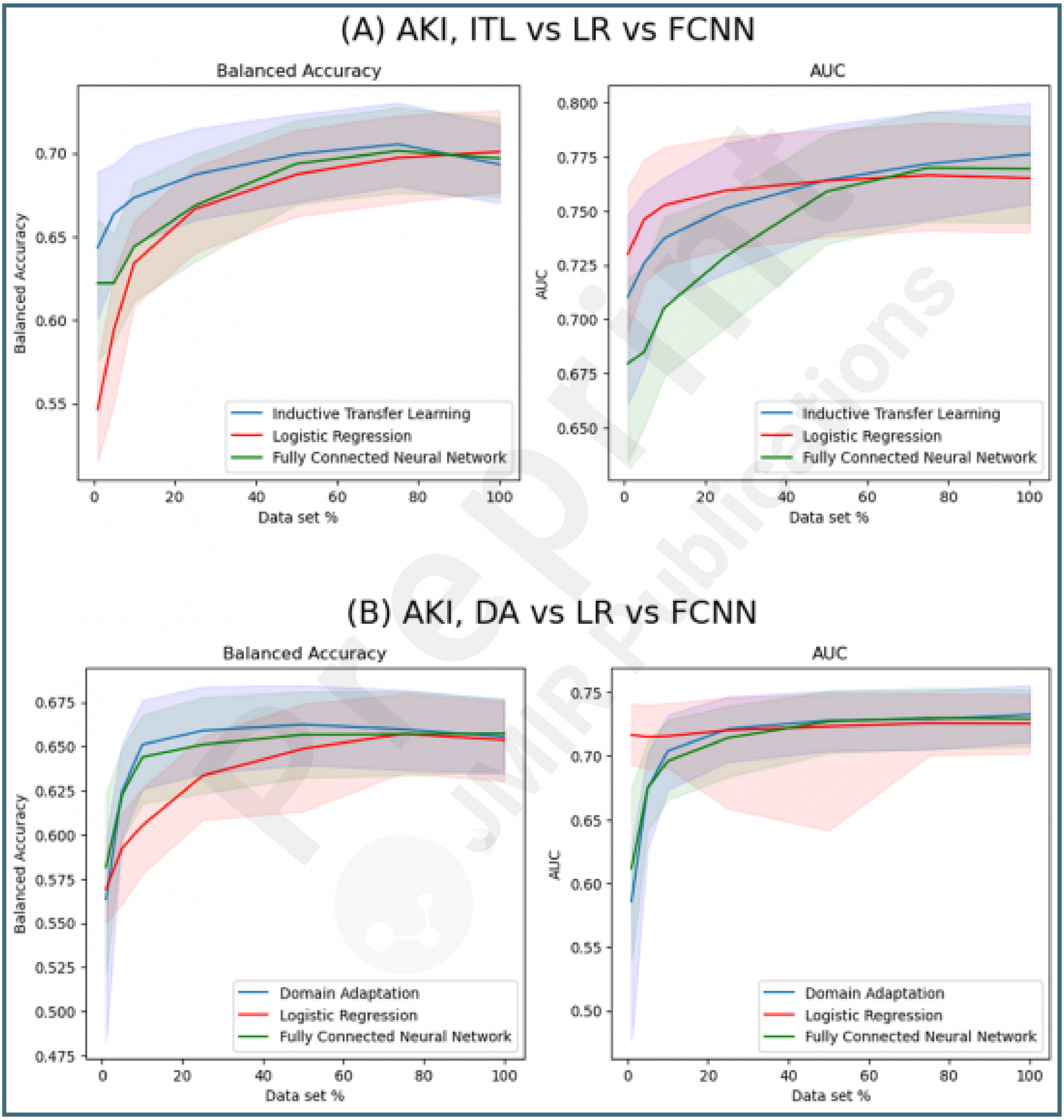
Patient cohort flowchart.



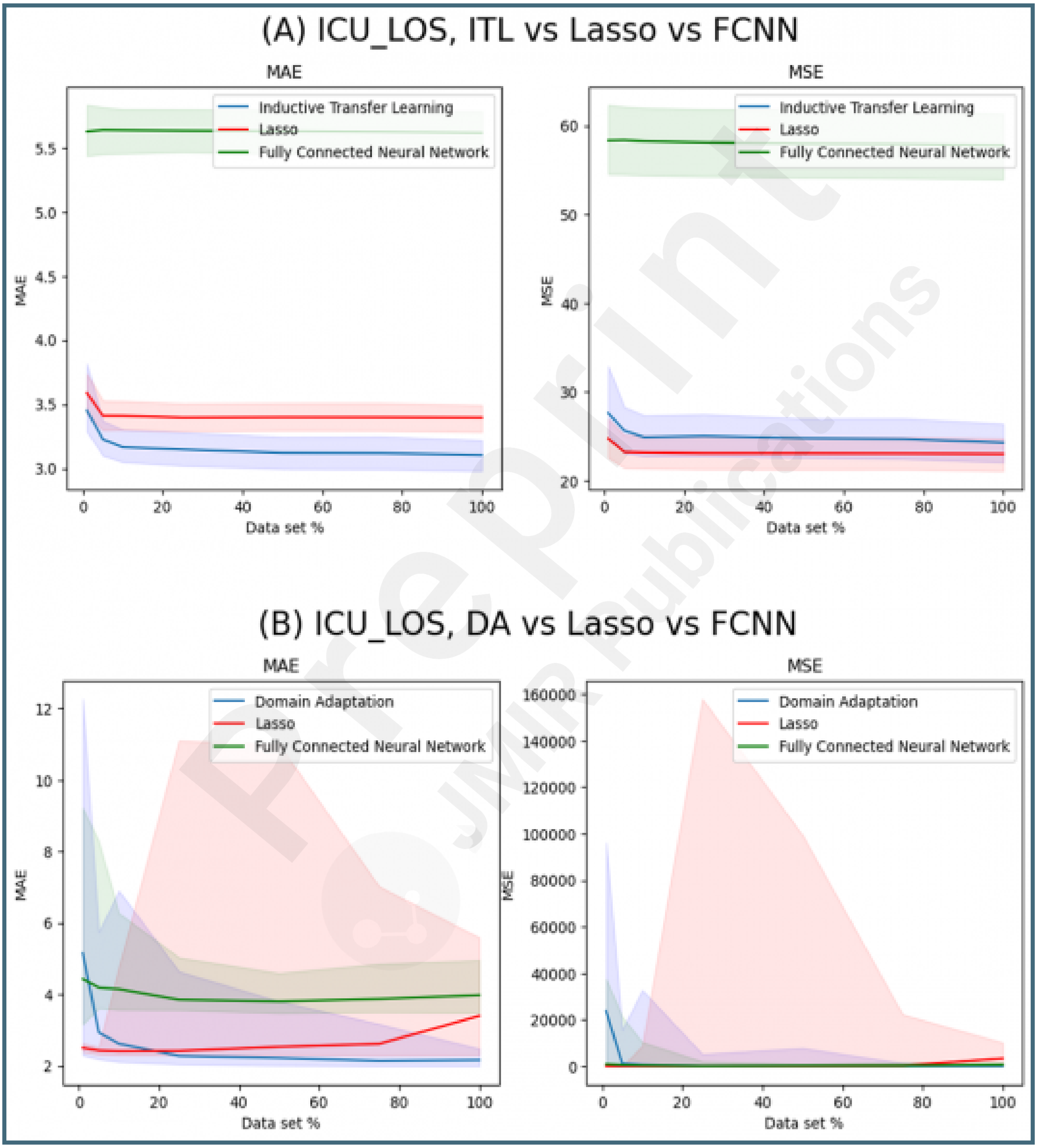
30-day mortality prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% confidence intervals. ITL: inductive transfer learning; DA: domain adaptation; LR: logistic regression; FCNN: fully connected neural network; AUC: area under the receiver operating characteristic curve.



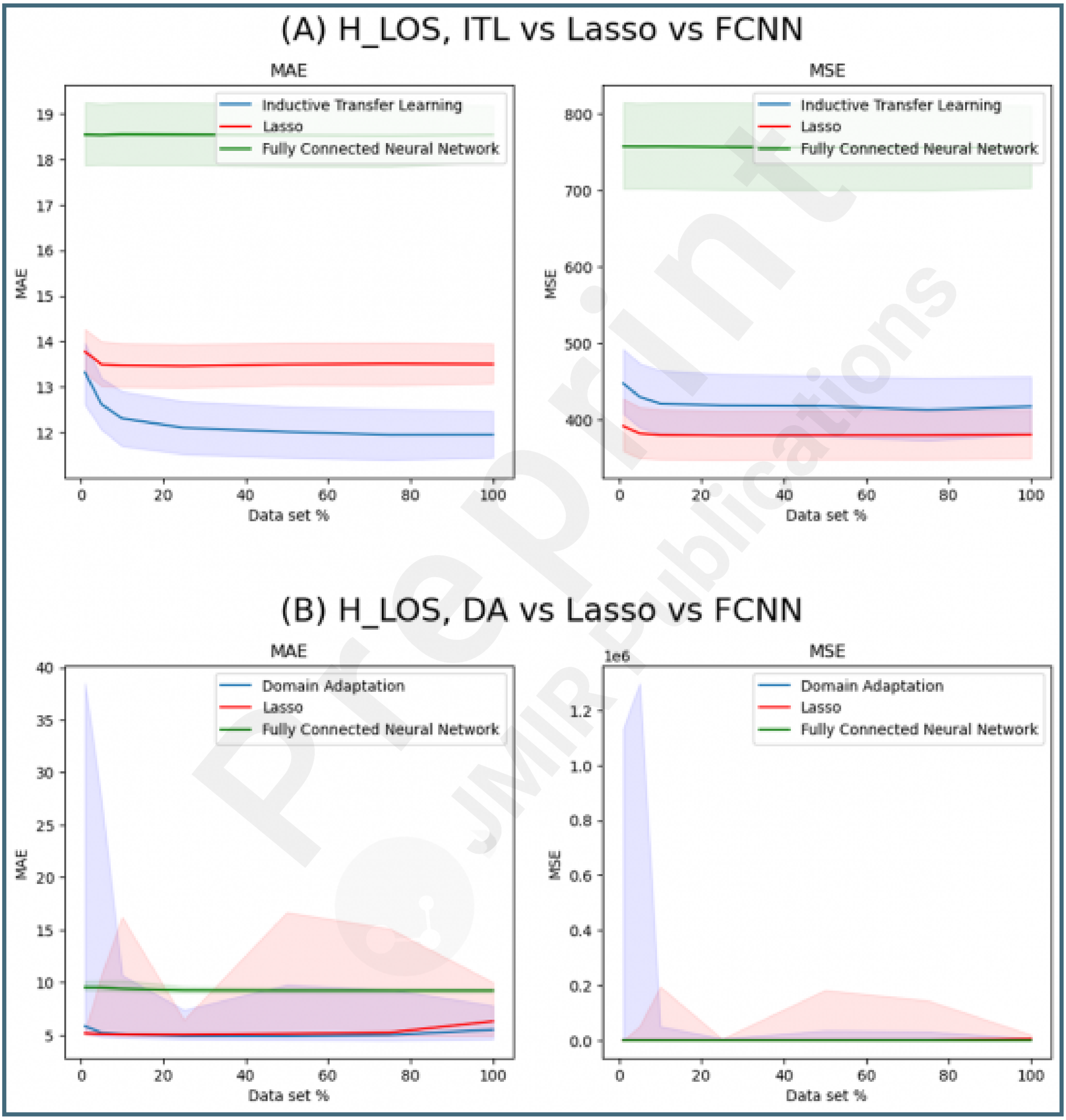
AKI prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% confidence intervals. ITL: inductive transfer learning; DA: domain adaptation; LR: logistic regression; FCNN: fully connected neural network; AKI: acute kidney injury; AUC: area under the receiver operating characteristic curve.



ICU_LOS prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% confidence intervals. ITL: inductive transfer learning; DA: domain adaptation; FCNN: fully connected neural network; ICU_LOS: intensive care unit length of stay; MSE: mean squared error; MAE: mean absolute error.



H_LOS prediction performances of the (A) ITL and (B) DA models in comparison with those of the baseline models across a range of data subsets representing varying levels of data scarcity. The solid lines are the medians and the shaded areas are the 95% confidence intervals. ITL: inductive transfer learning; DA: domain adaptation; FCNN: fully connected neural network; H_LOS: hospital length of stay; MSE: mean squared error; MAE: mean absolute error.



Multimedia Appendixes

30-day mortality prediction performances of DA and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/fb210d372b83b1fe2065241a8d5c9991.docx

AKI prediction performances of DA and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/c0f3afa80f5c485a088122ba0146fcca.docx

ICU_LOS prediction performances of ITL, DA, and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/5ee374c1680bdc73c0779edcffe6b0c0.docx

H_LOS prediction performances of ITL, DA, and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/3b2fb09df54d68245b0884b90457de02.docx

30-day mortality prediction performances of ITL and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/5385bbc8ec2488aa15746b9c4e7ee7d2.docx

AKI prediction performances of ITL and baseline models across all data subsets.
URL: http://asset.jmir.pub/assets/1ae6c804185676f808a638b92f058a30.docx