UNIVERSITY OF CALIFORNIA SAN DIEGO

Generalizable Risk Predictive Deep Learning Models

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

by

Fatemeh Amrollahi

Committee in charge:

Professor Shamim Nemati, Chair
Professor Gabriel Wardi, Co-Chair
Professor Tsung-Ting Kuo
Professor Lucila Ohno-Machado
Professor Debashis Sahoo

2023

The Dissertation of Fatemeh Amrollahi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To Mom & Dad

EPIGRAPH

True ease in writing comes from art, not chance,
As those move easiest who have learn'd to dance.
'T is not enough to no harshness gives offence,—
The sound must seem an echo to the sense.

Alexander Pope

You write with ease to show your breeding,
But easy writing's curst hard reading.

Richard Brinsley Sheridan

Writing, at its best, is a lonely life. Organizations for writers palliate the writer's loneliness, but I doubt if they improve his writing. He grows in public stature as he sheds his loneliness and often his work deteriorates. For he does his work alone and if he is a good enough writer he must face eternity, or the lack of it, each day.

Ernest Hemingway

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My PhD journey has been both tough and fulfilling. Looking back, I feel so grateful for all the people who supported, guided, and inspired me along the way.

Above all, I am profoundly grateful to my mother and father. Though I missed my mother, her enduring encouragement to chase my dreams and commit to my studies has been a guiding light in my life. Her steadfast faith in my potential has been the driving force behind my journey. Equally, I am deeply thankful to my father, whose countless sacrifices and teachings about perseverance and dedication have laid the foundation for my ambitions. Mom and Dad, this accomplishment is as much a testament to your love and guidance as it is to my efforts.

I would like to extend my heartfelt thanks to Professor Shamim Nemati, my advisor. Your immense knowledge, patience, and consistent guidance have been instrumental in shaping my research and my perspective. Your mentorship went beyond academics, teaching me life skills that I will carry with me always.

I also want to extend my gratitude to Dr. Wardi. His medical knowledge and assistance have been crucial in helping me maintain my health and focus during this demanding phase. His support was invaluable in this journey.

A special mention must also be made for my colleague, Dr. Supreeth Shashikumar. Your invaluable insights, encouragement, and constructive criticism were pivotal in refining my work. Your dedication to my progress and your attention to detail have left an indelible mark on my academic journey.

To all the others who have contributed directly or indirectly to my journey, your role has been cherished and will be remembered. This thesis is not just a reflection of my work but a tapestry of the guidance, love, and support I have received from all of you.

Chapter 2, is a reprint of the material as it appears in Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. Sci Rep 12, 8380 (2022).Amrollahi, F., Shashikumar, S.P., Holder, A.L.,Nemati,S. The dissertation author was the primary investigator and first author of this paper.

VITA

2014	Bachelor of Computer Engineering, University of Amirkabir, Tehran
2018	Master of Computer Science, Emory University
2019–2023	Research Assistant, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

F. Amrollahi, S. P. Shashikumar, A. L. Holder, and S. Nemati, "Leveraging clinical data across healthcare institutions for continual learning of predictive risk models," Sci. Rep., vol. 12, no. 1, Art. no. 1, May 2022, doi: 10.1038/s41598-022-12497-7.

F. Amrollahi, S. P. Shashikumar, A. Meier, L. Ohno-Machado, S. Nemati, and G. Wardi, "Inclusion of social determinants of health improves sepsis readmission prediction models," J. Am. Med. Inform. Assoc., vol. 29, no. 7, pp. 1263–1270, Jul. 2022, doi: 10.1093/jamia/ocac060.

F. Amrollahi,S. P. Shashikumar, H. Yhdego, A. Nayebnazar,N. Yung,G. Wardi,S. Nemati, "Predicting Hospital Readmission among Patients with Sepsis using Clinical and Wearable Data". medRxiv [Preprint]. 2023 Apr 11:2023.04.10.23288368. doi: 10.1101/2023.04.10.23288368. PMID: 37090521; PMCID: PMC10120792.

F. Amrollahi, S. P. Shashikumar, P.Kathiravelu,A. Sharma, S. Nemati, "AIDEx - An Open-source Platform for Real-Time Forecasting Sepsis and A Case Study on Taking ML Algorithms to Production", 2020 42nd Annual International Conference of the IEEE

F. Amrollahi, S. P. Shashikumar, F. Razmi, S. Nemati,"Contextual embeddings from clinical notes improves prediction of sepsis", AMIA Annual Symposium Proceedings 2020, 197

FIELDS OF STUDY

Major Field: Bio-informatics and System Biology (Biomedical Informaics)

ABSTRACT OF THE DISSERTATION

Generalizable Risk Predictive Deep Learning Models

by

Fatemeh Amrollahi

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2023

Professor Shamim Nemati, Chair
Professor Gabriel Wardi, Co-Chair

The broad adoption of Electronic Health Records (EHRs) accelerated the development and usage of Machine learning (ML) and Deep learning (DL) algorithms in clinical settings. The potential uses of ML and DL algorithms to augment clinical decision-making in domains such as forecasting disease onset and progression, predicting response to treatments, and optimization of treatment protocols are growing. While most existing ML/DL models are trained on single-centered data, multi-center datasets are becoming increasingly available. However, curation of such datasets is often time-consuming and lags behind shifts in disease prevalence and changes in workflow practices, which are known to cause data distribution shifts and degradation in

ML/DL model performance.

In addition, data privacy concerns and patient confidentiality regulations continue to pose a major barrier to multicenter EHR data access. In this work, we developed algorithms to enable DL models to transfer their knowledge across institutional boundaries and learn from new episodes of patient care without forgetting previously learned patterns. We validated and compared our methods in the context of early prediction of sepsis using data across four geographically distinct healthcare systems. We explore several methods to enhance the generalizability of DL models. We focus on three areas: Continual Learning, Federated Learning, and Generative Adversarial Networks (GANs), introducing new algorithms within each area and comparing their performance against state-of-the-art models. We have validated and compared these methods in one of the most challenging tasks for biomedical researchers: predicting the onset of sepsis in intensive care units.

Introduction

The broad utilization of artificial intelligence, especially its role in enhancing industrial efficiency, customer experiences, and revenue generation over recent years, has increased interest in its potential uses within healthcare domain. Deep learning, a subset of Machine learning tools, that uses neural networks with connected non-linear layers to analyze various forms of data, has become particularly prominent in various clinical domains such as disease diagnosis, patient prognosis, decision-making support, and treatment suggestions [1–6]. For these deep learning models to be widely accepted in clinical settings, they need to be adaptable to different care settings and prioritize patient data privacy. A significant portion of the existing ML-model applications rely on data either from a single hospital or multiple hospitals within a unified healthcare system standards where medical protocols are largely uniform. A notable challenge in this approach is that models developed using data from a single healthcare entity often suffer from lack of generalizability. This is due to varying factors such as regional demographics, medical equipment, electronic health record systems, data collection frequencies, and inconsistencies in clinical procedures, including the manner in which diseases are coded and defined. To gain wide clinical adoption, deep learning-based clinical models have to be generalizable and portable, and ensure the privacy of patients whose data are used for model training and evaluations. To make ML models more generalizable, one approach is to train ML models on data from large and diverse patient populations, similar to industry ML and DL applications. This approach requires data sharing and aggregation across institutional boundaries, which raises issues of patient privacy and consent.

A recent external evaluation of popular predictive ML models highlighted challenges

related to models' generalizability, especially when data distributions shift or when there are changes in the types of patient demographics [7]. One method to enhance the external validity of deep learning systems is to adjust or "fine-tune" these models for each new healthcare environment, a process known as Transfer Learning. However, this could lead to multiple variations of the same algorithm being used across different care settings. This multiplicity poses regulatory issues around change managements and scientific concern about generating consistent, generalizable knowledge and insights. Hence, it's crucial to develop learning algorithms and models that can utilize data from diverse patient groups while preserving privacy. Additionally, these models should come with clear change control strategies to ensure they perform consistently and safely across diverse scenarios.

Several methods can be employed to enhance the broad applicability or generalizability of ML models, and here,For this study, we will explore three of them in greater details:

Continual Learning (also known as Lifelong Learning): This approach involves training models to learn continuously over time. As new data becomes available, the model incorporates this data into its existing knowledge without forgetting previous information. This is particularly useful in dynamic environments where data patterns can change over time.

Federated Learning (also known as Distributed Learning): Instead of centralizing data to train models, federated learning allows for data to remain at its source (e.g., individual hospitals or devices). Models are trained locally, and only model updates or weights are shared and aggregated. This approach helps in dealing with data privacy concerns and can pull insights from diverse data sources without actually moving the raw data.

Generative Adversarial Network (also known as GAN): These sophisticated models are trained to generate new data samples that resemble the original data. GANs consist of two neural networks that are trained in tandem. The goal of GANs is to produce synthetic data samples that mirror the characteristics of the original data. GANs can generate synthetic Electronic Health Records (EHR) datasets, reducing the need to transfer real patient data and ensuring that deep learning models maintain their knowledge without compromising patient

privacy.

We validated and compared these methods in the context of the most challenging tasks for biomedical researchers, predicting the onset of sepsis in intensive care units where information overload poses cognitive burdens on the ability of bedside caregivers to integrate risk factors across diverse sources [8].

Chapter 1

Introduction to Sepsis and Predictive Risk Monitoring

1.1 Why generalizeable sepsis prediction models are needed?

Sepsis is a life-threatening condition that arises when the body's over react to infection.Body's response to an infection could injure its own tissues and organs known as septic shock [9]. Instead of being localized to a particular area, the body's immune response to infection could becomes overactive, releasing a large number of chemicals into the bloodstream this causes widespread inflammation. Untreated sepsis can lead to a cascade of changes that can damage multiple organ systems, causing them to fail and eventually death.

Sepsis has been recognized as one of the most ancient and challenging conditions in the medical world. It manifests as a severe organ malfunction resulting from an imbalanced response of the body to an infection [9]. In terms of its pathophysiology, when an infection strikes the human system, it induces a multifaceted and extended response. This includes both proinflammatory actions, which target the removal of the invaders, and anti-inflammatory mechanisms, which aim at infection clearance, tissue repair, but can also inadvertently lead to organ harm and secondary infections [10].

Additionally, the exaggerated inflammation results in impaired tissue oxygenation and as a result of which organ damage occurs. Sepsis is a major public health concern accounting for

more than $20 billion (5.2%) of total US hospital costs in 2011 [11]. The inpatient sepsis-related costs for Medicare beneficiaries were projected at $41.5 billion in 2018. Over 1.7 million US patients are affected by sepsis annually, with roughly 275 thousands progress to septic shock and death [12, 13]. Although sepsis might not be as publicly recognized as conditions like heart attacks, 6% of all US hospital patients are diagnosed with sepsis, in contrast to 2.5% diagnosed with heart attacks. Almost 35% of hospital deaths are due to sepsis, which is remarkably higher than the mortality rate of heart attacks, ranges between 2.7-9.6% [11].

Since 2004, the international alliance of critical care professionals and the Surviving Sepsis Campaign (SSC) have been addressing the inconsistency in sepsis treatment strategies by advocating evidence-based "sepsis care bundles" [14]. These bundles consolidate the findings of multiple studies, all pointing to enhanced sepsis outcomes when timely interventions using broad-spectrum antibiotics, IV fluids, and vasopressors are administered [15–18].

The most recent recommendation from the SSC is a 1-hr bundle that in addition to obtaining diagnostic tests like cultures and lactate levels, prescribes standard treatment with broad spectrum antibiotics, IV fluid, and vasoactive drugs if necessary, all within an hour of a sepsis diagnosis [19]. Despite the existence of effective treatment protocols, accurately identifying sepsis early, before obvious clinical signs, is one of the most important needs for modern medicine to be addressed.

A recent investigation by Seymour et al. [20] indicated that with every hour of delay in administering antibiotics to septic patients, the mortality risk increase by 4-8%. So early detection of sepsis in hospital settings is crucial. It's estimated that timely identification and treatment of septic patients in US hospitals could lead to reduced deaths, shorter hospital stays, and an approximate saving of $1.5 billion [21]. The main aim of this research is to utilize data from Electronic Health Records (EHRs) to create a generalizable deep learning model that facilitates the accurate recognition and identification of sepsis onset across diverse healthcare settings.

1.1.1 Prior work on predicting sepsis

Prior studies have shown that early identification of sepsis and initiation of treatment could improve the sepsis outcomes significantly. Early prediction and timely management can not only drastically reduce death rates but also diminish the associated long-term complications often suffered by sepsis survivors [8]. Therefore, developing mechanisms for the early prediction of sepsis is not only crucial for clinical practice but also it's a major step forward in enhancing healthcare delivery and patient outcomes. In the last decade, several data-driven approaches for predicting sepsis in the ICU have been presented. Many approaches selectively compare with simple clinical scores, such as SIRS, NEWS or MEWS [22, 23]. However, none of these scores are intended as specific, continuously-evaluated risk scores for sepsis. Nemati et al. [24] used a modified Weibull-Cox model on a combination of low-resolution Electronic Health Record (EHR) data and high-resolution vital signs time series data to predict onset of sepsis four hours in advance with an AUC of 0.85. Futoma et al. [25] applied multitask Guassian process to RNNs to predict sepsis, septic shock and in-hospital mortality. Authors used the publicly available MIMIC-III [cite mimic iii] dataset and tried to predict sepsis 0 to 12 hours ahead. They compared their models with several alternative machine learning-based models for predicting sepsis.

Desautels and colleagues [26] attempted to predict sepsis onset in ICU patients using minimal EHR data (i.e., vital signs and limited lab tests) and a combination of different machine learning models. Their model was able to predict sepsis 4-12 hours prior to clinical recognition with an area under the receiver operating characteristic curve (AUC) of 0.83-0.85 using data from two hospitals. Shashikumar et al. [27] utilized Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks to predict onset of sepsis using data from 2 separate hospitals. Henry et al. [28] introduced a targeted real-time warning score (TREWScore) to predict septic shock, using real-time EHR data. Septic shock occurs when sepsis leads to low blood pressure that persists despite treatment with intravenous fluids.

However, external generalizability and validation of these models under data distribution

shift remains unclear and these models are likely to produce sub-optimal performance. A recent independent external validation of a widely used machine learning-based sepsis prediction risk score highlighted the issue of these models generalizability in the presence of data distribution shift and changes in the population case-mix [29].

In this research, we developed multiple appraiches and algorithm to enhance generalizability of our deep learning predictive models, focusing on forecasting the onset of sepsis using Electronic Health Record (EHR) data available in Intensive Care Units (ICUs). We used data from four geographically distinct hospitals, with unique data distributions.

Chapter 2

Leveraging Clinical Data Across Healthcare Institutions for Continual Lifelong Learning of Predictive Risk Models

2.1 Introduction

The digital transformation of healthcare has brought a new era of medical innovation, promising more precise, accessible, and personalized patient care. As AI integrates deeper into clinical environments, the application of advanced ML-based algorithms, particularly deep learning models, has showcased potential in analyzing complex patterns avaiable in EHR and medical data records. However, a critical limitation with these models is their difficulty in generalizing knowledge across diverse healthcare systems, given the inherent variability in patient demographics, equipment, and clinical practices. Such challenges often result in models customized for specific datasets, making them less effective when confronted with new, unseen data. Addressing these challenges, continual learning emerges as a promising solution. Continual learning, by allowing models to learn progressively and adjust to new information without losing previously acquired knowledge, offers the possibility of closing the gap in generalizability for healthcare applications. This approach moves us nearer to achieving universally applicable digital healthcare solutions.

Continual learning offers a robust methodology for sequentially learning multiple tasks

while preserving efficiency on tasks previously learned. Continual learning stands as a promising solution to the challenge of learning successive tasks without compromising the proficiency achieved on prior tasks. This addresses the challenge of 'catastrophic forgetting,' where deep learning models tend to forget previously acquired knowledge when exposed to new data. By leveraging knowledge from past experiences, continual learning enhances the model's ability to rapidly adapt to new tasks. This capability is especially suitable in the medical domain, where clinical data are divided across various institutions. Nonetheless, a unified understanding across diverse patient groups is critical.

The ability to incrementally learn from experiences in a continual learning framework is ideal for medical applications where clinical data are siloed across different institutions, and yet generalizability is desirable across patient cohorts. We hypothesized that continual learning might be advantageous in clinical settings where a model is expected to generalize across multiple healthcare systems, and data is intended to remain local to each healthcare institution

We introduce WUPERR, our state-of-the-art continual learning approach designed to learn sequentially from data streams without forgetting prior knowdgle. Instead of transferring raw data, WUPERR emphasizes on sharing model weight uncertainties and learned representations across institutions. This facilitates learning from diverse datasets while maintaining data privacy and granted each institution's control over its data. We show that propagating model weight uncertainty and learned representations across healthcare sites (as opposed to the raw data) allows WUPERR to learn from diverse datasets while preserving privacy of data. The innovation in WUPERR is inspired by cutting-edge advancements in the lifelong learning realm. Specifically, it integrates the idea of Elastic Weight Consolidation (EWC) [30] and Episodic Representation Replay (ERR) [31] to continuously refine our predictive models as they encounter new patient cohorts from varied geographical locales. WUPERR combines Episodic Representation Replay (ERR) and Weight Uncertainty Propagation (WUP) drived from EWC to enable continual learning of tasks while mitigating the problem of catastrophic forgetting. The goal of WUPERR is to minimize the drop in performance on older tasks when

the model is trained on a new task (i.e., a new hospital). WUPERR attempts to achieve this goal through consolidation of network parameters important to model performance on prior tasks and episodic experience replay (by maintaining sample data representations during prior training and periodically revisiting those examples at new task). To achieve privacy, WUPERR replaces raw patient-level features with hidden representations learned via a neural network (i.e., lower level of the neural network), thus obviating the need for moving protected health information outside institutional boundaries.

In this chapter, we delve deep into the details of WUPERR, elucidating its functionality and nuances as a representative of continual learning approaches in clinical domain. We further, evaluate it against established and widely-recognized baseline algorithms in this domain. Through this comprehensive exploration, we aim to provide a clear understanding of WUPERR's unique attributes and its standing in the realm of continual learning within the healthcare domain.

2.2 Methods

2.2.1 Study Populations

In this study, we employed a cohort comprising 104,000 adult patients admitted to Intensive Care Units (ICUs) across four geographically distinct healthcare institutions: UC San Diego Health, Emory University Hospital, the Beth Israel Deaconess Medical Center, and Grady Hospital, referred to as Hospital-A through Hospital-D respectively. Our analysis rigorously adhered to all relevant guidelines and regulations. The Institutional Review Boards (IRBs) of UC San Diego (IRB#191098), Emory University/Grady Hospital (IRB#110675), and the Beth Israel Deaconess Medical Center (IRB#0403000206) approved the use of the de-identified data for this study. Furthermore, the requirement for informed patient consent was waived by these IRBs, as the Health Insurance Portability and Accountability Act (HIPAA) privacy regulations exempt the use of de-identified retrospective data from this prerequisite.

Patients in the cohort, all aged 18 years or above, were tracked throughout their ICU stay

until the onset of their first sepsis episode or until their transfer out of the ICU. We used the definitions provided by the Third International Consensus Definitions for Sepsis (Sepsis-3). The diagnostic criteria for sepsis onset in our study were twofold: (1) an indication of an infection (2) evidence of acute organ dysfunction through at least 2 point changes in SOFA score.

Evidence of Acute Organ Dysfunction: This was identified by a notable increase in the Sequential Organ Failure Assessment (SOFA) score by a minimum of two points. We took into consideration organ dysfunction symptoms that manifested between 48 hours prior and 24 hours post the time an infection was suspected.

Suspicious of Infection: Clinical signs an infection was defined by specific actions – namely, the initiation of a blood culture draw and the subsequent administration of intravenous (IV) antibiotics for a period of at least three consecutive days. The sequencing of these actions is important: if the blood culture was drawn before the antibiotics were prescribed, the antibiotics had to be administered within the next 72 hours. Conversely, if the antibiotics were prescribed first, the blood culture had to be initiated within the subsequent 24 hours.

The onset time of sepsis was identified then as the earliest time of suspicious of infection and evidence of acute organ dysfunction. We focus on predicting sepsis hourly, starting from the fourth hour post ICU admission.This timeframe ensures adequate initial patient assessment and stabilization, as well as a comprehensive data collection scope for predictive purposes. We excluded patients diagnosed with sepsis before our prediction onset, those lacking heart rate or blood pressure measurements prior to prediction initiation, and individuals with ICU stays exceeding 21 days.

2.2.2 Data Preprocessing

A total of 40 clinical variables were extracted across the four hospitals (see Table 2.1). Additionally, for every vital signs and laboratory variable, their local trends (slope of change) and the time since the variable was last measured (TSLM) were recorded, resulting in a total of 108 features (the same set of variables have been used in a previously published study [32]).

The patient characteristics of all the four cohorts have been tabulated in Table 2.3. Figure 2.1 illustrates the differences in race ethnicity across the four datasets. Table 2.2 shows the missingness percentage acccross the four cohorts. All continuous variables are reported as medians with 25% and 75% interquartile ranges (IQRs). Binary variables are reported as percentages. All vital signs and laboratory variables were organized into 1-hour and 1-day non-overlapping time series bins to accommodate for different sampling frequencies of available data for the sepsis cohort. All the variables with sampling frequencies higher than once every hour (or day) were uniformly resampled into 1-hour (or 1-day) time bins, by taking the median values if multiple measurements were available. Variables were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). Mean imputation was used to replace all remaining missing values (mainly at the start of each record).

Table 2.1. List of Clinical variables used for hourly prediction of the risk of developing sepsis

Variable	Measurement Unit
Vital Signs (Dynamical Features)	
Heart rate	beats/minutes
Mean Arterial Pressure	mmHg
Pulse oximetry	%
Diastolic BP	mmHg
Temperature	degC
Respiration rate	Breaths/minutes
Systolic BP	mmHg
End tidal Co2	mmHg
Systolic BP	mmHg
End tidal Co2	mmHg
Laboratory values (Dynamical Features)	
Excess bicarbonate	mmol/L
Serum Glucose	mg/dL
Bicarbonate	mmol/L
Lactic acid	md/dL
Fraction of inspired Oxygen	%
Magnesium	mmol/dL
pH	-
Phosphate	mg/dL
Partial Pressure of CO2	mmHg
Potassium	mmol/L
Oxygen Saturation	%

Table 2.1. List of Clinical variables used for hourly prediction of the risk of developing sepsis

Variable	Measurement Unit
Total Bilirubin	mg/dL
Aspartate transaminase	IU/L
Troponin I	ng/mL
Blood Urea Nitrogen	mg/dL
Hematocrit	%
Alkaline Phosphate	IU/L
Hemoglobin	g/dL
Calcium	mg/dL
Partial thromboplastin time	Seconds
Chloride	mmol/L
White blood count	$10^3/L$
Creatinine	mg/dL
Fibrinogen	mg/dL
Bilirubin direct	mg/dL
Platelets	$10^3/L$
Demographics	
Age	Years
Pre ICU Stay	Hours
Gender	Male/Female
ICU Length of Stay	Hours
Careunits	Medical/Surgical ICU

Table 2.2. Summary of missingness percentage of the variables across four cohorts considered in this study.

Labs/Vitals	Hospital-A %missing	Hospital-B %missing	Hospital-C %missing	Hospital-D %missing
Heart rate	11.90	13.64	1.95	9.94
Pulse oximetry	13.10	15.61	2.04	17.90
Temperature	63.37	66.74	5.80	46.53
Systolic BP	36.79	15.38	2.25	10.69
Mean arterial pressure	37.00	16.41	2.02	11.57
Diastolic BP	36.80	15.39	2.26	10.69
Respiration rate	13.04	21.47	1.96	11.34
End tidal CO_2	88.36	93.62	90.51	97.32
Excess bicarbonate	96.99	99.81	93.12	95.63
Bicarbonate	97.00	99.86	91.34	99.68
Fraction of inspired Oxygen	80.41	97.85	41.66	90.98
pH	96.98	97.88	88.40	95.57
Partial pressure of CO_2 from arterial blood	97.00	98.87	79.46	95.63
Oxygen saturation from arterial blood	97.00	98.20	80.55	99.72
Aspartate transaminase	97.82	98.17	69.40	96.55
Blood Urea Nitrogen	92.44	94.38	84.22	92.76
Alkaline phosphatase	97.81	98.17	88.69	96.55
Calcium	92.54	93.07	83.99	92.72
Chloride	92.55	99.44	83.65	92.75
Creatinine	92.50	94.37	88.15	92.74
Bilirubin direct	99.26	99.77	95.07	96.55
Serum Glucose	92.48	78.32	42.25	82.45
Lactic acid	98.59	98.23	91.34	98.06
Magnesium	94.16	95.23	76.42	95.07
Phosphate	94.60	97.07	87.74	95.73
Potassium	92.01	92.37	43.51	92.68
Total Bilirubin	97.84	98.16	73.09	96.55
Troponin I	98.63	97.90	72.17	98.67
Hematocrit	92.45	94.19	43.68	92.44
Hemoglobin	92.44	94.14	89.73	89.69
Partial Thromboplastin Time	96.09	99.06	89.95	99.37
White Blood Cell count	92.45	94.69	89.08	92.82
Fibrinogen	99.40	99.51	99.81	99.55
Platelets	92.46	94.63	84.59	92.84

Table 2.3. Summary of patient characteristics of the four cohorts considered in this study(including University of California San Diego Health, Emory University Hospital, MIMIC IV, and Atlanta's Grady Hospital, referred to as Hospital-A, Hospital-B, Hospital-C, and Hospital-D respectively)

	Hospital-A		Hospital-B		Hospital-C		Hospital-D	
patient characteristics	Sepsis	Non-Sepsis	Sepsis	Non-Sepsis	Sepsis	Non-Sepsis	Sepsis	Non-Sepsis
Patients (#)	13,208	3,825	37,899	7,913	28,892	7,916	2,999	992
Male	789	2463	20313	4305	16000	4439	2303	779
Race								
Caucasian	7,013	1,897	20,882	3,885	19540	5,288	789	227
African American	977	295	13,871	3,331	3340	843	1946	688
Asian	713	241	947	166	896	221	36	7
Other	4,505	1,410	2,199	531	5,116	1564	228	70
Age [IQR]	60.2 [46.8 71.3]	60.7 [48.1 72.2]	62 [50 72]	62 [50 72]	64 [52 76]	64 [52 75]	55 [37 66]	57 [42 67]
T_sepsis (hrs)[IQR]	-	22.6 [11.2 52.6]	-	26.3 [10.3 67.3]	-	9 [1 36]	-	37 [12 85]
ICU-LOS(hrs)[IQR]	45.6 [25.8 80.9]	188.7 [93.4 381.3]	45.2 [25.5 76.5]	176.3 [87.2 353.1]	36 [60 24]	95 [49 201]	56.8 [32.0 94.4]	256.7 [127.3 449.1]
SOFA	3 [1 5]	7 [4 10]	2 [1 5]	7 [4 10]	2 [1 6]	6 [3-9]	3 [1 6]	9 [6 12]
Mortality	357	845	827	1282	2059	1546	214	233

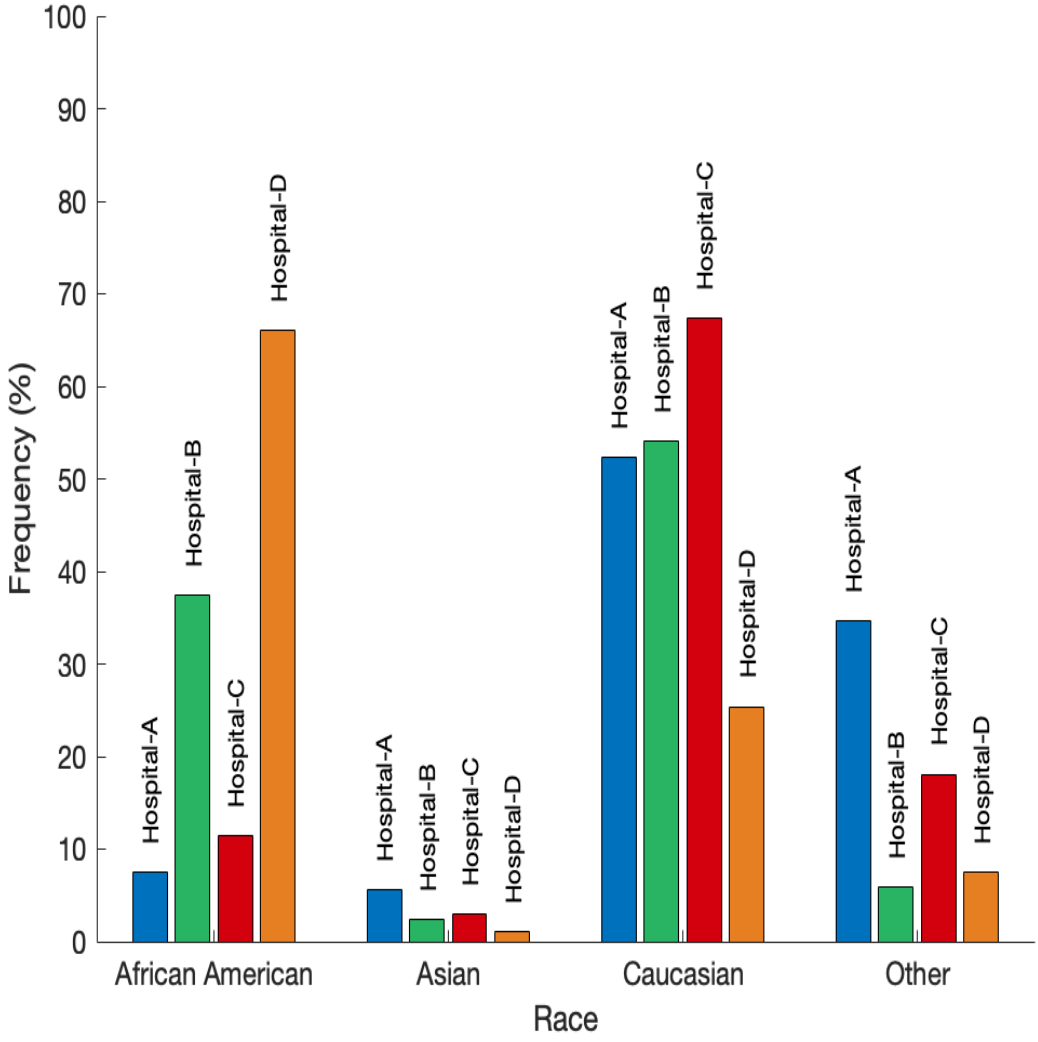

Figure 2.1. Distribution of race/ethnicity per cohort

2.2.3 Continual Learning

Continual learning, often referred to as ”lifelong learning” or ”incremental learning”, embodies the concept of training machine learning models to learn progressively over time. Instead of a one-time training on a fixed dataset, these models are designed to learn new information without forgetting the previously acquired knowledge. This capability is crucial in clinical setting where data streams are non-stationary over the time, arriving sequentially, and where retraining a model from scratch every time is inefficient. Continual learning is inspired by the human brain's ability to acquire and integrate new knowledge over time without necessarily forgetting past experiences. The brain possesses several mechanisms for lifelong learning, which researchers aim to mimic through continual learning approaches.

One of the most remarkable attributes of the human brain is its neuroplasticity, the ability to reorganize itself by creating new neural connections throughout life [33]. This plasticity allows humans to adapt to new experiences, learn new information, and even recover from brain injuries. In continual learning for DL models, this adaptability is replicated by allowing models to modify their weights in response to new data, while we need to ensure important parameters remain relevant to prior tasks as they encounter different tasks or data distributions.

The strengthening and weakening of synapses is believed to underlie the storage of memories in the brain. This process of synaptic consolidation plays a role in how the brain balances between retaining old memories and forming new ones [34]. In the context of artificial neural networks, certain weights (analogous to synapses) in the network are protected or ”consolidated” to prevent them from drastic changes, ensuring that prior knowledge is not easily overwritten by new information. While the brain's propensity to forget may seem like a limitation, there's growing evidence suggesting that active forgetting is an adaptive process. It ensures that only pertinent information is retained, while less relevant or potentially misleading information is discarded. In continual learning, this idea is captured by the selective updating of model parameters. Not all parts of the model are adjusted for every new task; instead, the model

”chooses” which parts to update, replicating the brain’s selective memory retention.

The human brain has a hierarchical and modular organization. Different regions of the brain specialize in different functions, and information processing can occur in a layered manner, from basic perceptions to complex cognitions [35]. Drawing from this, some continual learning methods propose modular architectures for neural networks, where different sub-networks or modules are trained for different tasks, thereby reducing interference between tasks.In addition, the brain often ”replays” experiences, especially during rest or sleep, which is believed to play a role in memory consolidation. In the DL context, the replay strategy involves periodically revisiting old data while training on new tasks, ensuring that the model doesn’t forget its earlier training.

while artificial neural networks and the human brain are fundamentally different in their architectures and mechanisms, the principles governing memory retention, adaptability, and learning in the brain provide invaluable inspiration for developing robust continual learning strategies in machine learning.

Continual learning approaches can be grouped in to four main types including:

- **Rehearsal Methods:**Rehearsal methods involve storing some or all of the past data and ”replaying” them when training on new data. This approach tries to blend old and new information, ensuring that the model does not drift too far from its previous knowledge while still accommodating the new data. The primary advantage of replay is its straightforward nature and the direct way it tackles catastrophic forgetting. However, the challenge lies in managing the storage requirements especially when dealing with large datasets, and privacy of patient data by moving them. Some variations of replay, like ”pseudo-replay”, generate synthetic samples instead of storing actual data, helping to alleviate these concerns. Replay-based Methods in continual learning have various innovative implementations. Mnih et al. [36] utlized experience replay for stabilizing deep reinforcement learning. Experience Replay is perhaps the most straightforward incarnation

of the replay paradigm. Originally popularized in the context of Deep Q-learning for reinforcement learning, it involves maintaining a memory buffer that stores previous data samples. During training, instead of solely using new data, random samples are drawn from this buffer and mixed with the new data for a more balanced training process.The primary motivation behind Experience Replay is to break the temporal correlations of incoming data, providing a more i.i.d (independently and identically distributed) experience to the model. This helps mitigate catastrophic forgetting as the neural network gets reminders of its past knowledge. However, managing the size of the replay buffer becomes crucial as one needs to strike a balance between the diversity of samples and storage limitations. The other approach are Generative Replay methods. Generative Replay leverages generative models, like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), to generate synthetic samples of past data. Instead of storing actual past examples, a generative model is trained to capture the distribution of previous datasets. When new tasks come in, this generative model can produce synthetic examples from past tasks which are then combined with the new data for training. The strength of Generative Replay lies in its ability to produce a potentially infinite amount of data from previous tasks without the need for extensive storage. However, the quality of the replay and the efficacy of the continual learning process are dependent on the quality of the generative model. Poorly generated samples can negatively affect the learning process [37].The last well-known replay approach is Pseudo-Rehearsal [38].Pseudo-Rehearsal is a variant of replay where, instead of using generative models, a separate neural network is employed to produce ”pseudo-samples.” This network, trained on earlier tasks, generates outputs for random inputs, effectively creating synthetic examples that capture the characteristics of past tasks.Pseudo-Rehearsal can be seen as a form of distillation, where the knowledge from one network is transferred to another. It has the benefit of not requiring the explicit storage of old data samples and not being dependent on sophisticated generative models. However, the quality and diversity of pseudo-samples can vary, and their appropriateness

for the task can sometimes be a concern [38].

- **Regularization-based Methods:**These methods add constraints to the neural network's loss function to prevent drastic changes in the learned weights when training on new tasks. The aim is to ensure that the updated weights after learning a new task remain close to the original weights from prior tasks. One prominent technique in this category is "Elastic Weight Consolidation" (EWC), which penalizes changes to network weights that are important for previous tasks [30]. This method allows for the retention of old knowledge while making room for the new. The challenge here is determining the importance of weights and ensuring that the regularization does not overly constrain the learning of new tasks. EWC introduces a regularization term to the loss function that penalizes changes to network weights that were important for previously learned tasks. EWC achieves the importance of weights by computing the Fisher Information Matrix, a measure that identifies which weights in the network are crucial for the tasks the model has already learned.

  Synaptic Intelligence (SI) is another regularization based method that, like EWC, focuses on identifying important weights in the network [39]. However, instead of the Fisher Information, SI computes a surrogate measure for the importance of each weight by considering the contribution of the weight to the overall loss over its entire learning trajectory. When training on a new task, SI imposes a penalty on weights with higher importance values, ensuring they don't change drastically. This method is particularly notable for its efficiency, as it requires fewer computations than EWC.

  Cuong et al. [40] introduce Variational Continual Learning (VCL) as a regularization based method which employs a Bayesian neural network approach for continual learning. Instead of single-point estimates for each weight, VCL maintains a distribution over the weights.This method addresses catastrophic forgetting by incorporating the uncertainty inherent in neural network training. As new data comes in, the posterior distribution from

the previous task becomes the prior for the next task. With this approach, VCL provides a principled method to update the neural network's weights while considering the uncertainty associated with them.

- **Dynamic Architectural Methods:** Within this approach instead of fitting all the knowledge into a fixed architecture, these methods dynamically modify the network structure. For instance, as new tasks arrive, new neurons or layers might be added to the network. Progressive Neural Networks is a notable technique in this category, where each new task, or data distribution gets a new set of neurons, and these neurons are interconnected. These approaches address the catastrophic forgetting through avoiding overwritten to previous knowledge. The primary challenge with dynamic methods is the potential growth in the network's size, making it less efficient over time. Andrei et al. [41] introduce ProgNNs, the idea of adding new columns or sub-networks for each new task. Essentially ProgNNs preserves the weights and structure of a network as they are when a new task arrives. Instead of retraining the network, a new column is added to netwrok. New connections allow the new column to access the representations of previous ones, but not vice versa, ensuring that the new task does not interfere with the already learned tasks. The downside of ProgNNs is the potential for the model size to grow substantially with each new task. DENs [42] take a more flexible approach compared to ProgNNs by allowing for selective expansion of the network. DEN is trained in an online manner trough performing selective retraining, dynamically expansion of network with only the necessary number of units. DEN evaluates which parts of the network to expand, rather than adding an entirely new column of neurons. DEN freezes the previously learned weights, then identifies which parts of the network are crucial for the new task and, based on this analysis, decides where to expand the network. Furthermore, DEN has a pruning process to remove less important neurons to maintain efficiency. In [43] as new tasks arrive, the model undergoes architecture optimization to determine the most effective structure for accommodating

the new data or task without forgetting the old. In this work, authors employs Neural Architecture Search (NAS) techniques to dynamically adapt the architecture of the network for sequential learning. Instead of manual or heuristic-based decisions on how the architecture should change, NAS automates the process, searching for the optimal structure that can cater to both new and old tasks. Though promising, a significant challenge here is the computational expense associated with NAS to explore search space of all possible structures.

- **Context-based Methods:** These methods, through determining the context (or task) at hand, adjust the network parameters to achieve good performance across various tasks. Here, different tasks can have dedicated pathways or shared representations in the network. Context-based methods operate under the assumption that providing a model with task-specific context improve interference between tasks and eventually avoid catastrophic forgetting. Context-dependent gating [44], uses a mechanism where certain parts of the neural network are selectively activated based on the arrival task. This is achieved through gating mechanisms which ensure that only a specific subset of the network's units are active. These gates can be controlled using external signals indicating the current task or learned in an end-to-end fashion. By ensuring only a subset of neurons are active for any given task, this approach reduces interference between tasks. The advantage of gating mechanism is that by limiting the active parts of the network, it minimizes the overlap and thus reduces the interference between tasks. However, determining the optimal gating mechanism is challenging.

  Task-Embedded Control Networks (TECNs) is another context based approach where a task embedding, is used to modulate the behavior of a neural network [45]. In TECNs, first, a task descriptor is computed. This descriptor, or embedding, is then used throughout the network to modulate its behavior. This can be done through mechanisms like feature-wise transformations or by influencing gating mechanisms. By embedding task-specific

information directly into the network's operations, TECNs ensure that the model is aware of the current task context, which can help in preserving previously learned tasks and efficiently learning new ones.

Similarly, in Conditional Neural Processes (CNPs) [46] they utilize context to make predictions. CNPs take a set of context points and an input, and they produce a distribution over possible outputs. Effectively they are combining ideas from meta-learning and Bayesian neural networks. The network here is structured to use context points, adapting its behavior based on the provided context. For a given task, there is a set of context points and a set of target points. The context points provide information about the task, and the target points are what you want to predict. The context points are fed into an encoder, which produces a fixed-size representation. This context representation aims to capture the underlying structure of the task based on the context points. The context representation, along with target inputs, is then passed to a decoder. The decoder produces a distribution over possible target outputs. This distribution explicitly models the uncertainty in the prediction. During training, the context points are randomly sampled from the available data, teaching the CNP to make predictions based on varying amounts of context. This equips the CNP to handle few-shot scenarios where only limited context might be available. In the realm of continual learning, they can be seen as a way to use task-specific context to adjust the model's behavior rapidly and reduce interference between tasks.

These strategies constitute key components of the extensive spectrum of continual learning methodologies, each characterized by unique advantages and limitations. The selection of a suitable approach is typically contingent upon the particularities of the problem at hand, the characteristics of the incoming data stream, and the extent of computational resources.

Continual learning offers a robust methodology for sequentially mastering multiple tasks or data stream while preserving efficiency on tasks or data previously learned without the need for transferring data outside of institutional boundaries. Despite the need for

robust continual learning algorithms in clinical settings, applications of such methods to clinical predictive modeling remain scarce [47]. Here we consider a clinically significant problem involving prediction of sepsis in critically ill patients. Using data across four sepsis cohorts, we developed and validated a continual learning framework for sequentially training predictive models that maintain clinically acceptable performance across all cohorts while preserving patient data privacy.

In our research, we employ continual learning to empower deep predictive models to learn from data streams incrementally and sequentially, sourced from diverse institutions. Our objective is to achieve this without the models losing their proficiency on data from previous hospitals, while simultaneously safeguarding patient data confidentiality. Here, we introduce the **WUPERR** algorithm, a novel continual learning strategy. We are inspired by the idea of replay-based and regularization-based techniques to continual learning. The subsequent sections will delve into the WUPERR algorithm, detailing its algorithm and assessing its performance against multiple baseline methodologies in the domain of sepsis prediction.

2.2.4 WUPERR

WUPERR combines Episodic Representation Replay (ERR) and Weight Uncertainty Propagation (WUP) to enable continual learning of tasks while mitigating the problem of catastrophic forgetting. The goal of WUPERR is to minimize the drop in performance on older tasks when the model is trained on a new task (i.e., a new hospital). WUPERR attempts to achieve this goal through consolidation of network parameters important to model prediction on prior tasks (via a targeted weight regularization scheme) and episodic experience replay (by maintaining sample data representations encountered during prior training and periodically revisiting those examples during re-training). Figure 2.2 shows the schematic diagram of the WUPERR algorithm.

Via the WUPERR algorithm and after we fine-tuning the model on Hospital-A data, we freeze the shallow (input) layer of the network. By passing the input data through the first network layer, WUPERR obtains a data representation that we leverage towards continuous learning (these representations are no longer considered protected health information). To retain previous knowledge we replay the Hospital-A data representations while fine-tuning the model on Hospital-B data using a variant of weight uncertainty propagation. Similarly, as we go to the Hospital-C we carry forward the weight uncertainties for regularization purposes and fine-tune the deeper layers of the model using the representation of Hospital-A and Hospital-B, as well as data from Hospital-C data.

Let N, J, K be the number of parameters of the neural network, the number of training epochs, and the total number of tasks, respectively. At training time of task k, the loss $L(j;\theta)$ calculated at epoch j is as follows:

$$L(j,\theta) = L_{CE}(j;\theta) + \frac{\gamma}{2}\sum_{n=1}^{N} I_n^k(j-1)(\theta_n^k(j-1) - \theta_n^{k-1})^2 \tag{2.1}$$

where $L_{CE}(j;\theta)$ corresponds to the cross-entropy classification loss, $\theta_n^k(j-1)$ corresponds to the n-th parameter of the neural network from the previous epoch, $I_n^k(j-1)$ is an approximation of Fisher information (inverse of uncertainty) associated with parameter θ_n during task k and epoch $j-1$. The approximate Fisher information corresponding to parameter θ_n during task k and epoch j is computed as follows:

$$I_n^k(j) = \beta * I_n^k(j-1) + (1-\beta)(\frac{\partial L(j;\theta)}{\partial \theta_n^k})^2 \tag{2.2}$$

Note that the magnitude of the gradient corresponds to the degree of steepness of the loss surface around a point in the parameter space, which in-turn provides a measure of information gain. For task $k(k = 2,\ldots,K)$, I_n^k is initialized as $max(I_n^1,...,I_n^{k-1})$.

Bayesian Optimization is used to optimize regularization parameter (equation (2.1)) and

uncertainty estimation moving average parameter (equation (2.2)).

Note that, after task 1 (Hospital-A), parameters corresponding to the first layer of the neural network are frozen (a.k.a representation layer). Additionally, after completion of training on each Task k, the hidden representations (h_1^k; output from the first layer of neural network) corresponding to a random sample of patients from Hospital-k are stored. From Task 2 onwards, we fine-tune the neural network (except for the first layer) with data from the new patient cohort (Hospital-k) and hidden representations stored from previous tasks.

2.2.5 Baseline models

The performance of the WUPERR algorithm was compared against four baseline models, listed below:

- **site-specific training:** In this approach, we trained the model in isolation at each hospital site wherein a new model is trained on each task independently.
- **Transfer learning:** Transfer learning is a technique where a model developed for a particular task or data distribution is repurposed on a second related task or data. It is most effective when the features learned from the first task are relevant to the ongoing tasks. Transfer learning assumes that the source and target tasks are derived from the same feature space, as a result of which transferring knowledge from prior tasks might accelerate the learning procedure on new tasks and thereby improve model performance.However, one of the challenges that arises in transfer learning is the issue of domain shift, where the data distribution of the source domain is different from the target domain [48].To address this, transfer learning techniques often incorporate fine-tunings.
    - Transfer learning: In this approach, parameters of the neural network after training on task k-1, was transferred over to task k and were further fine-tuned using data from task k.

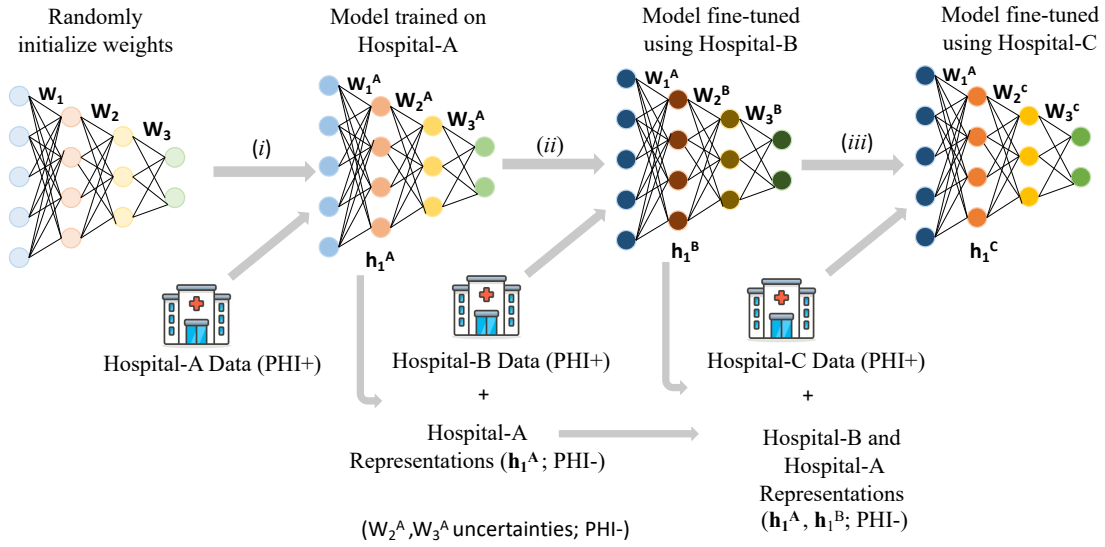

Figure 2.2. Schematic diagram of the WUPERR algorithm. The training starts with a randomly initialized set of weights, which are trained on the first task (e.g., prediction on Hospital-A data). In all subsequent learning tasks the input layer weights (W_1^A) are kept frozen. The optimal network parameters, the parameter uncertainties under task-A, and the set of representations from training cohort of Hospital-A ($\{h_1^A\}$) are then transferred to Hospital-B. The deeper layers of the model are fine-tuned to perform the second task (e.g., prediction on Hospital-B data) through replaying the representation of Hospital-A and Hospital-B data. Similarly, the optimal parameters and their uncertainty levels along with the Hospital-A and Hospital-B representations are transferred to Hospital-C to fine-tune the model on performing the third task. Note, at no time protected health information (PHI+) leaves the institutional boundaries of a given hospital.Finally, at the time of evaluation (on testing data) at a given task, the model is evaluated on all the hospital cohorts.

* Transfer learning-freeze:In this approach, the first layer of the neural network was frozen after training on task 1. Parameters of the neural network after training on task k-1, were transferred over to task k and were further fine-tuned (all layers except the first layer) using data from task k.

– **Elastic weight Consolidation (EWC) [30]:** EWC balances the trade-off between stability and plasticity to avoid forgetting in incremental task learning. EWC through protecting previously learned knowledge, allows the network to maintain existing knowledge (stability) and still have the ability to learn new information (plasticity). This balance is critical in achieving effective continual learning without suffering from catastrophic interference. EWC relies on regularization terms to avoid forgetting. EWC protects the neural network performance on old tasks by slowing down the learning process on selected weights and staying in a region corresponding to lower error for prior tasks while learning a new task.To identify weights that carry more information, EWC relies on a fisher information matrix. EWC is formalized mathematically by adding a regularization term to the loss function during training. Let a neural network has parameters θ, after fine-tuning on task 1 we have θ^*, then we want our neural network to learn task 2.EWC adds a quadratic penalty to the loss function $L_2(\theta)$ for new task, leading to the modified loss function 2.3:

  Here,λ is the a hyperparameter that controls the strength of the regularization,θ_i represents the individual parameters of the network, and $_i$ represents the diagonal elements of the Fisher Information Matrix of task 1, evaluated at θ^*. A large value of $_i$ indicates that the i-th parameter is very important for preserving the performance on task 1, and therefore, the penalty for changing it is high.

$$L(\theta) = L_2(\theta) + \frac{\lambda}{2}\Sigma_i F_i(\theta_i - \theta_i^*)^2 \tag{2.3}$$

– **Episodic Representation Reply (ERR):**In ERR, we use representations of data

from previous tasks in addition to data from the current to fine tune a model. I'm We observed the greatest changes in the network weights at the deeper layers, which may suggest that these layers are more important to learning a new task. consequently, it was observed that freezing the weights within the first network layer had little effect on the ability of the network to adapt to a new dataset. This enabled us to use the first layer (after training on Task 1) as an encoding network to obtain representations for the upper network layers. From Task 2 onwards, we used these input data representations at every new site, in conjunction with the representation of data from prior sites, to train the model. The latter (i.e., replaying data representations from prior tasks) enabled the network to remember the older tasks while learning from a new dataset.

2.3 Results

We evaluated the performance of the proposed continual learning algorithm for early prediction of onset of sepsis in hospitalized patients across four healthcare systems. A comparative study of WUPERR against several baseline models is shown in this section.

2.3.1 Evaluation metrics

To gauge the effectiveness of the continual learning models, we conducted a comparative analysis of each approach utilizing the Clinical Workflow aware AUC (C-AUC) as introduced by Shashikumar et al. [49], alongside metrics such as Positive Predictive Value (PPV) and Sensitivity. We proceed to detail modifications to the c-AUC algorithm and elucidate why substituting AUC with c-AUC is imperative in the context of this study.

Although, well-known performance metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (AUCpr) are

standard benchmarks for evaluating the efficacy of predictive models. However, in the context of healthcare we need to define a suitable and more specific quality metrics.

In healthcare setting two types of data available: longitudinal, where patient data is recorded over time as a sequence, and static, where a single-time-point snapshot of patient information is captured. Our research focuses on models that process longitudinal or time-sequenced patient data. The AUC and AUCpr metrics are enable to account the temporal dimension of the data, considering each data point independently of its place in time. To address this,Shashikumar et al. [49] introduce the Clinical Workflow Area Under the Curve (C-AUC), which has been conceptualized with two critical considerations in mind: firstly, it's metric for evaluating the Clinical Decision Support (CDS) system application and secondly, its accommodation for the sequential progression of patient data.

The first modification that they make to the AUC/AUCpr metric is that any predictions after a positive prediction has been made are ignored for a fixed duration (which they call as 'snooze duration').

The second alteration addresses the progressive nature of conditions like Sepsis, which doesn't manifest suddenly but develops over time, making it challenging to pinpoint a precise moment of onset. Such gradual progression is not acknowledged by traditional AUC/AUCpr metrics. These metrics assign a binary label, negative until four hours before the diagnosed onset of the disease, switching to positive in the four-hour window leading up to the disease's actual onset. This abrupt binary classification penalize predictive models. For instance, if a model predicts the onset of the disease earlier than four hours before the actual onset, this would be inaccurately recorded as a False Positive. Such predictions outside the designated four-hour predictive window are erroneously penalized. To mitigate this issue and reduce unfair penalization for early predictions, they propose a refinement in the computation of False Positives (FP), True Positives (TP), False Negatives (FN), and True Negatives (TN) by adjusting the criteria to accommodate predictions made

before the horizontal window.

Here,we tried to develop a generalizable DL model that can predict the onset of sepsis four hours in advance. So corresponds to the [49] the FP,TP,FN and TN is:

- **FP:** makes a positive prediction - M hours prior to the onset of sepsis (here we cosider M=12 hours) for septic patients or at any point of time for non-septic patients.
- **TP:** makes a positive prediction - within M hours prior to onset of sepsis
- **FN:** makes a negative prediction - between four hours prior to onset of sepsis
- **TN:** makes a negative prediction four hours prior to onset of sepsis for septic patients or at any point of time for non-septic patients.

In our research, we define a timeframe referred to as 'positive prediction duration', denoted by M,to be 12 hours. This duration is the time-frame within which the model is not penalized for positive prediction for the onset of sepsis.Therefore, any positive prediction made by the model within this 12-hour window leading up to the actual occurrence of sepsis is considered acceptable and not subject to penalty.However, it is important to note that the model will still incur a penalty if it fails to predict sepsis within the critical four-hour period before the onset of sepsis.

Finally, the proposed C-AUC in [49] takes into consideration both the first modification and second modification.

In our study, we fixed the snooze duration at 6 hours and positive prediction at 12 hours while computing C-AUC/C-AUCpr.

2.3.2 Evaluation setting

The WUPERR framework was used to train a model to sequentially predict the onset of sepsis (defined according to the Sepsis-3 consensus definitions for Sepsis and Septic Shock)

four hours in advance [50]. To investigate the impact of variations in data distributions on our model performance, we trained our model sequentially on over 104,000 patients belonging to four critical care centers with various underlying demographic characteristics. The model was first trained on the Hospital-A dataset (Task 1), followed by Hospital-B (Task 2), Hospital-C (Task 3) and Hospital-D (Task 4).

We compare the performance of WUPERR algorithm in terms of c-AUC, PPV, and Sensitvity with other five baseline continual learning methods for a fixed threshold of sensitivity equals to 0.8 at hospital-A. Figure 2.3 displays evaluation of continual learning models for early prediction of sepsis using PPV metric. 2.3a illustrates the positive predictive value (PPV) of four separate models trained at each site separately on all the other sites. In all cases, PPV was calculated at a fixed threshold, corresponding to 80% sensitivity at Hospital-A. For instance, the model trained at hospital-C (with PPV of 31%) performs poorly on hospital D (PPV of 24%). Figure 2.3b illustrates the model PPV on sequentially learning to predict the onset of Sepsis across four distinct hospitals using Transfer learning.Figure 2.3 (c-f) shows the same for Transfer-Learning-Freeze, Elastic Weight Consolidation (EWC), Episodic Representation Replay (ERR), and the proposed WUPERR method, respectively.Figure 2.3g shows PPV values on Hospitals A-C after continual learning on all four hospitals with site-Specific (orange), Transfer learning (red), EWC (green), ERR (purple) and WUPERR (blue). Similarly, Figure 2.4, and 2.5 illustrate the assessment of WUPERR performance on predicting onset of sepsis and other five base line continual learning models in terms of C-AUC, and Sensitivity respectively.

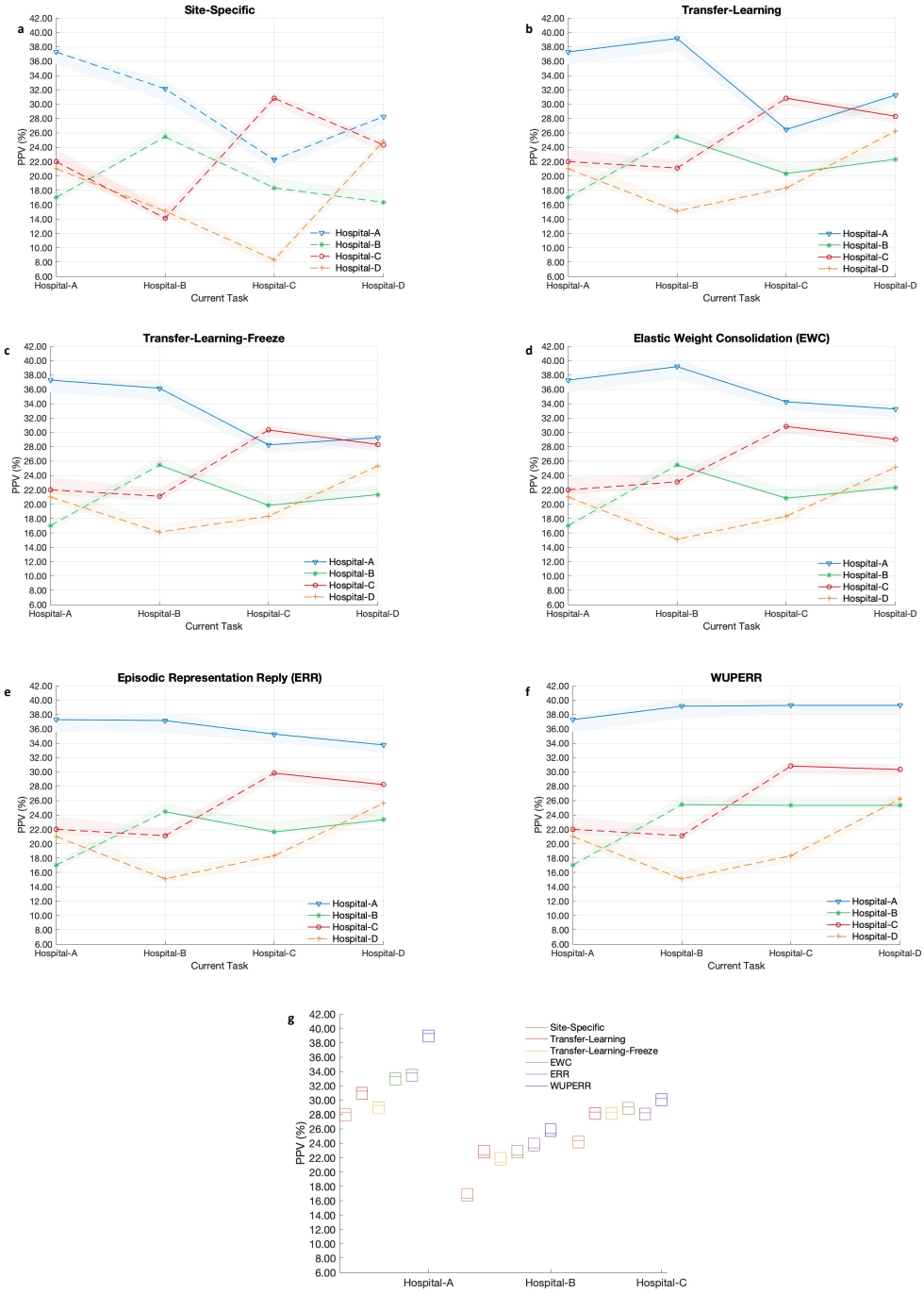

Figure 2.3. Evaluation of continual learning models for early predicting of onset of Sepsis. Evaluation of continual learning models for early prediction of sepsis using PPV metric.In all cases, PPV was calculated at a fixed threshold, corresponding to 80% sensitivity at Hospital-A.

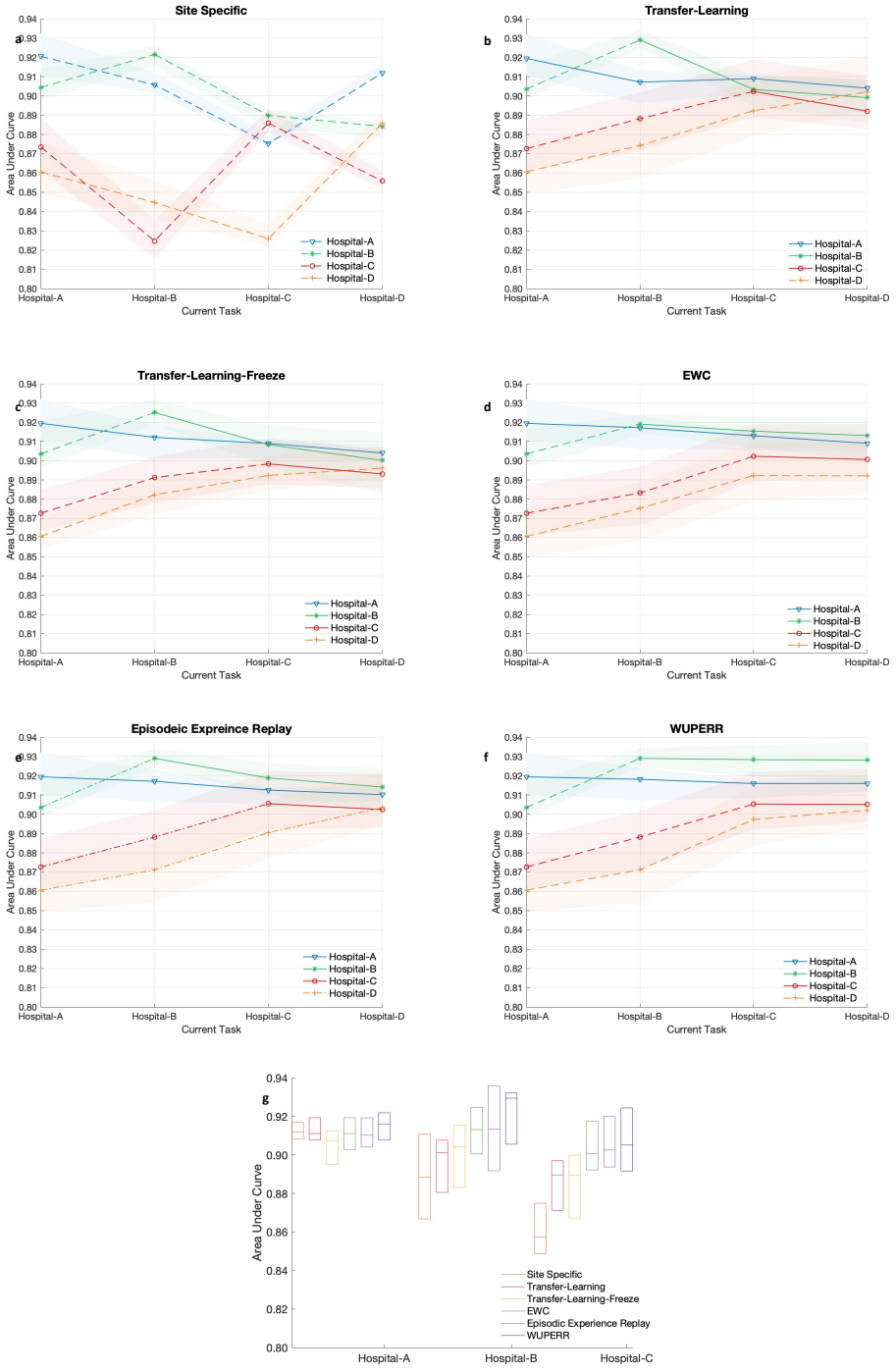

Figure 2.4. Evaluation of continual learning models for early predicting of sepsis measured using C-AUC metric.

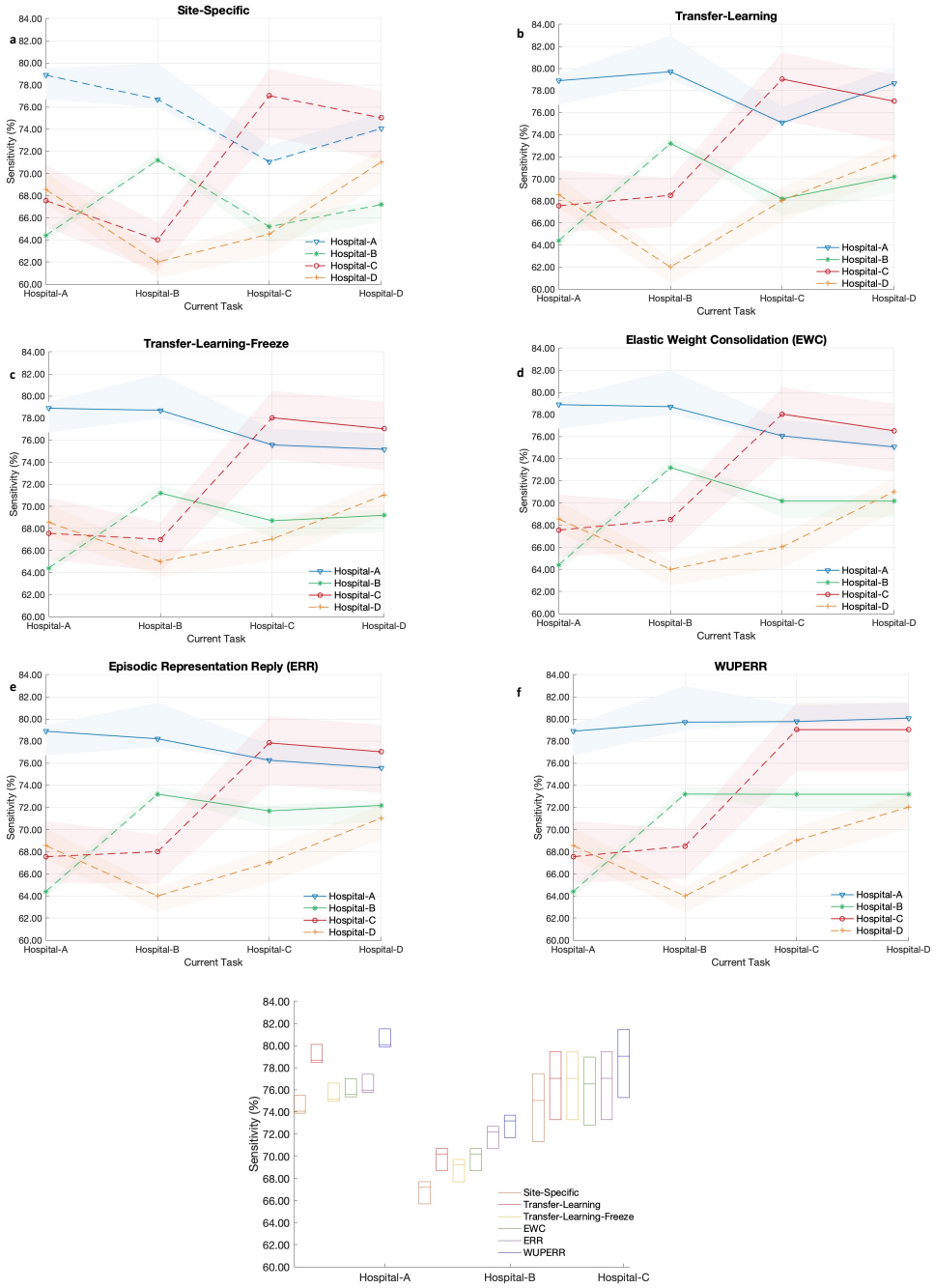

Figure 2.5. Evaluation of continual learning models for early predicting of sepsis. Same as the previous figure but here we summarize the model sensitivity at the fixed threshold.

Throughout the subsequent stages of our experimental analysis, we maintained the integrity of the context while streamlining our comparison process. We focused solely on benchmarking our algorithm against a baseline transfer learning method for clarity and conciseness.This approach allows for a more direct and focused analysis of our proposed method's performance.

Figure 2.6, panels a-c, show the performance of WUPERR on the four hospital datasets, where the model was trained on one cohort at a time and the performance is reported on testing data from all other cohorts (previous and subsequent cohorts). With the transfer learning approach, we observed that with the progression in training on new cohorts the model performance degenerated on previous cohorts. Whereas sequential training by WUPERR enabled the model to maintain comparable performance on older tasks. For example, at the end of Task 4 with transfer learning, AUC of the model on Task 2 was 0.90 [0.89-0.91], a drop from the AUC of 0.93 [0.92-0.94] when the model was trained on the data from Hospital-B (corresponding to task 2). In comparison, at the end of Task 4 with WUPERR, the model maintained its performance on Task 2 with an AUC of 0.93 [0.91-0.94]. Notably, we observed that the superiority of WUPERR over transfer learning grow as the number of subsequent training cohorts the model was exposed to increased (see Fig. 2.6, panel c, performance on Hospital-A at the end of training on hospital-D). Additionally, we observed that at the end of Task 4, the model trained with the WUPERR approach performed superior to transfer learning across all the Hospital cohorts (See Fig. 2.6b).

In Fig. 2.7 we compared the positive predictive value (PPV) of the model sequentially trained on four cohorts using the WUPERR approach versus the baseline transfer learning approach. A decision threshold corresponding to 80% sensitivity was chosen after completion of training on Task 1. This decision threshold was then used to measure positive predictive value (PPV) for all the remaining tasks. We observed that WUPERR

consistently outperformed the transfer learning approach across all the tasks (See Fig. 2.7a-c). For instance, with WUPERR the positive predictive value (PPV) for Hospital-A improved from 37.28 [35.57-37.69] after Task 1 to 39.27 [38.11-39.78] by the end of Task 4, whereas with transfer learning approach the positive predictive value (PPV) dropped to 31.28 [30.11-31.78] by the end of Task 4.

Additionally, WUPERR was able to maintain consistent sensitivity levels on the Hospital-A cohort while being sequentially trained on Tasks 2, 3, and 4 (79.70 [78.50-82.57], 79.76 [79.57-81.20], 80.06 [79.87-81-50], respectively). In comparison, the sensitivity level on the Hospital-A cohort dropped below 80% when the model was trained on Tasks 2, 3 and 4 in the case of transfer learning approach (See Fig. 2.7d). Similar patterns for sensitivity were observed for the other hospital cohorts.

Lastly, we evaluate the resilience of WUPERR to the sequence of datasets in which the model is trained. Figures 2.8,2.9, 2.10, and 2.11 display the performance assessment of our proposed WUPERR algorithm against standard Transfer-Learning-based continual learning approaches while the order of datasets is swapped. This is done using PPV and Sensitivity as metrics for the early prediction of Sepsis, under conditions where the sequence of hospital datasets is altered. From these observations, it is evident that WUPERR maintains its robustness regardless of the training sequence, and it consistently surpasses the Transfer-Learning approach.

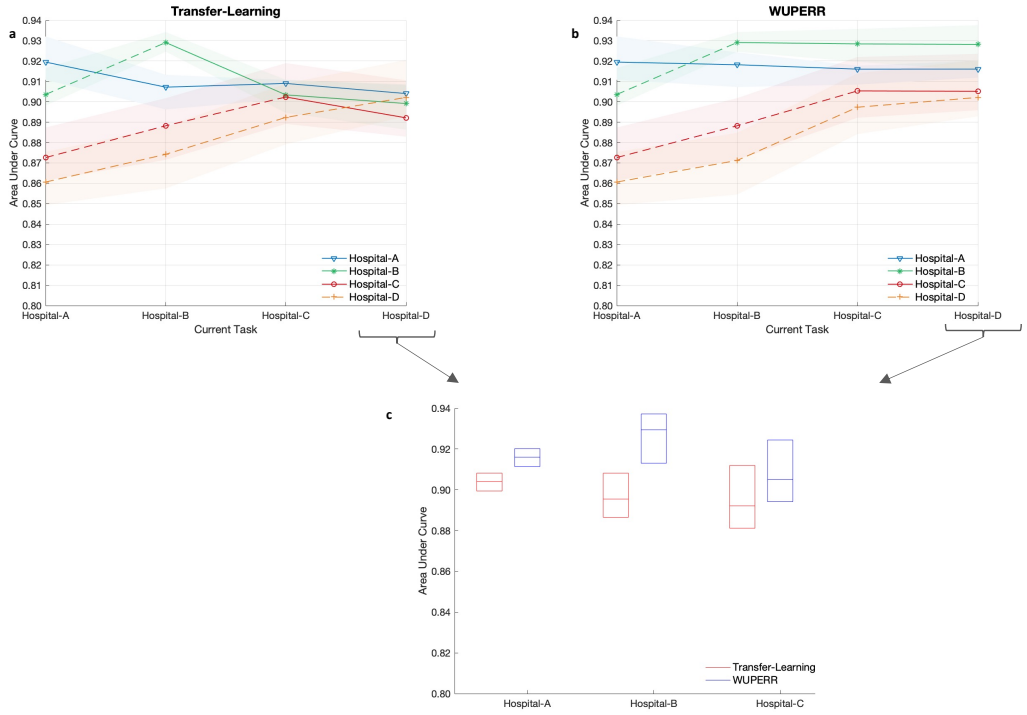

Figure 2.6. Evaluation of continual learning models for early predicting of onset of Sepsis, redmeasured using Area Under the Curve (C-AUC) metric. Panel (a) Illustrates C-AUC of a model (median[IQR]) trained using transfer learning. The model performance is reported (using different markers; see legend) across all the cohorts after sequential training on data from a given hospital on the x-axis. Panel (b) shows the C-AUC of the proposed WUPERR model, under the same experimental set-up as panel (a). redAt the time of evaluation (on testing data) at a given site, the model is evaluated on all the hospital cohorts. The solid line-style indicates that at the time of model evaluation (on testing data) at a given site, the model had already seen the training data from that site. For instance, since the model is first trained on Hospital-A data, the performance of the model on this dataset after continual learning on all subsequent hospitals is shown in solid line-style to signify that the model had already seen this patient cohort in the past. Panel (c) summarizes the model performance (median[IQR]) on Hospitals A-C after continual learning on all four hospitals with Transfer learning (red) and WUPERR (blue).

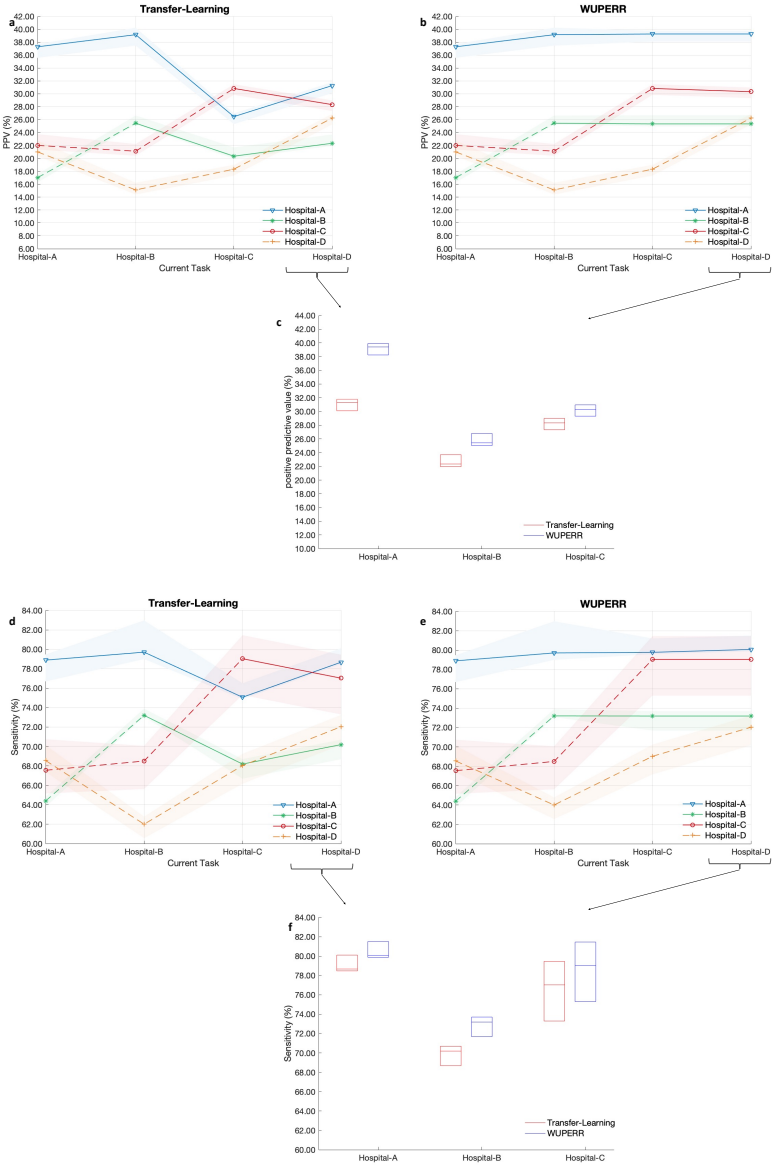

Figure 2.7. Evaluation of continual learning models for early predicting of onset of Sepsis,measured using positive predictive value (PPV) and sensitivity. Panel (a) Illustrates the PPV of a model (median[IQR]) trained using transfer learning (measured at fixed threshold of 0.41 corresponding to 80% sensitivity at Hospital-A after Task 1, for all folds and across all tasks). Panel (b) shows the PPV of the proposed WUPERR model, under the same experimental set-up as panel (a). Panel (c) summarizes the model performance (median[IQR]) on Hospitals A-C after continual learning on all four hospitals with Transfer learning (red) and WUPERR (blue). Panels (d-f) summarize the model sensitivity results under the same experimental protocol.

In all the aforementioned assessment strategies, we utilized data from Hospital-A to train the representation layer, which constitutes the initial layer of our neural network. Once this layer was trained, we fixed its parameters across subsequent processing tasks. Consequently, the quality of data from Hospital-A is pivotal for effectively training representation layer. Additionally, we evaluated the network's performance when Hospital-D was designated as the initial training task (see Fig.2.12 and Fig.2.13). The PPV for Hospital-C, after the model sequentially learned from Hospital-D, Hospital-C, Hospital-B, and finally Hospital-A, was found to be 23.50[22.89-23.73]. In contrast, the PPV for Hospital-B, when the learning sequence was Hospital-A, Hospital-B, Hospital-C, and then Hospital-D, yielded a different result of 25.34[24.83-26.59] (comparing Fig.2.12 and Fig.2.7).

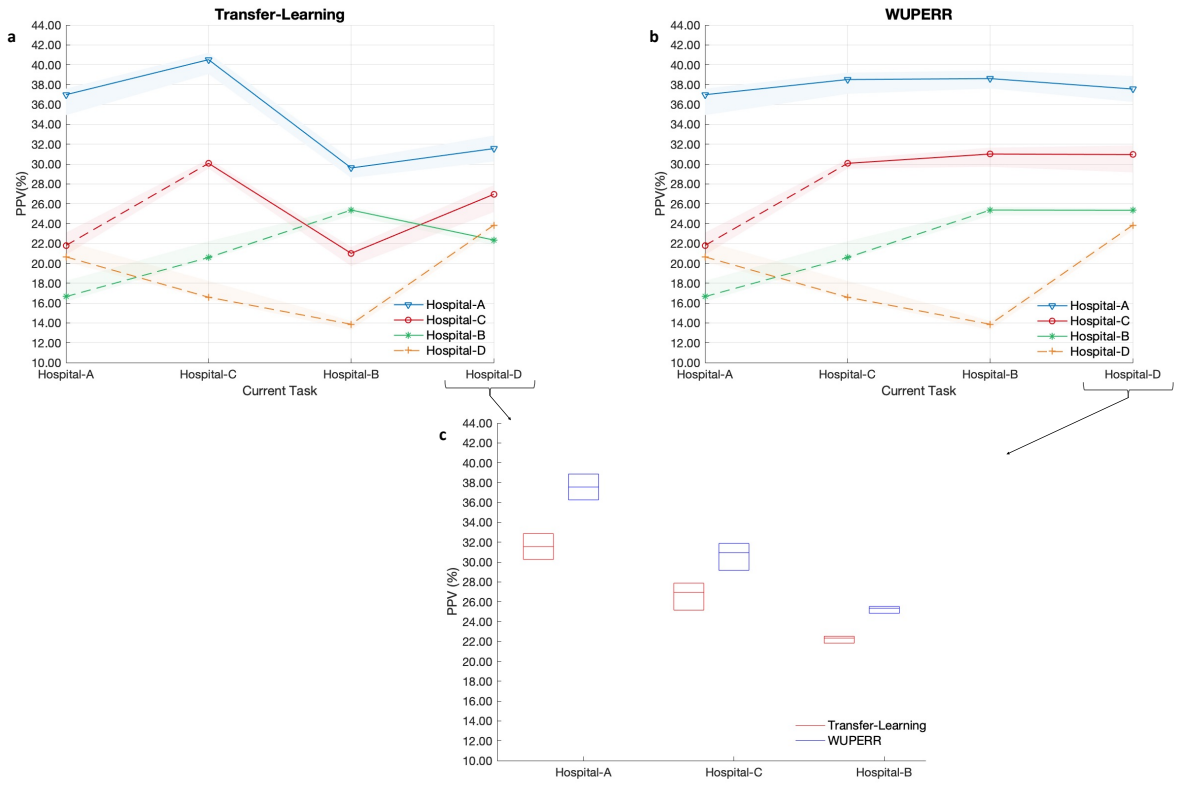

Figure 2.8. Evaluation of continual learning models for early predicting of onset of Sepsis. The performance measured using PPV metric while tasks reordered as Hospital-A, Hospital-C, Hospital-B, and Hospital-D.

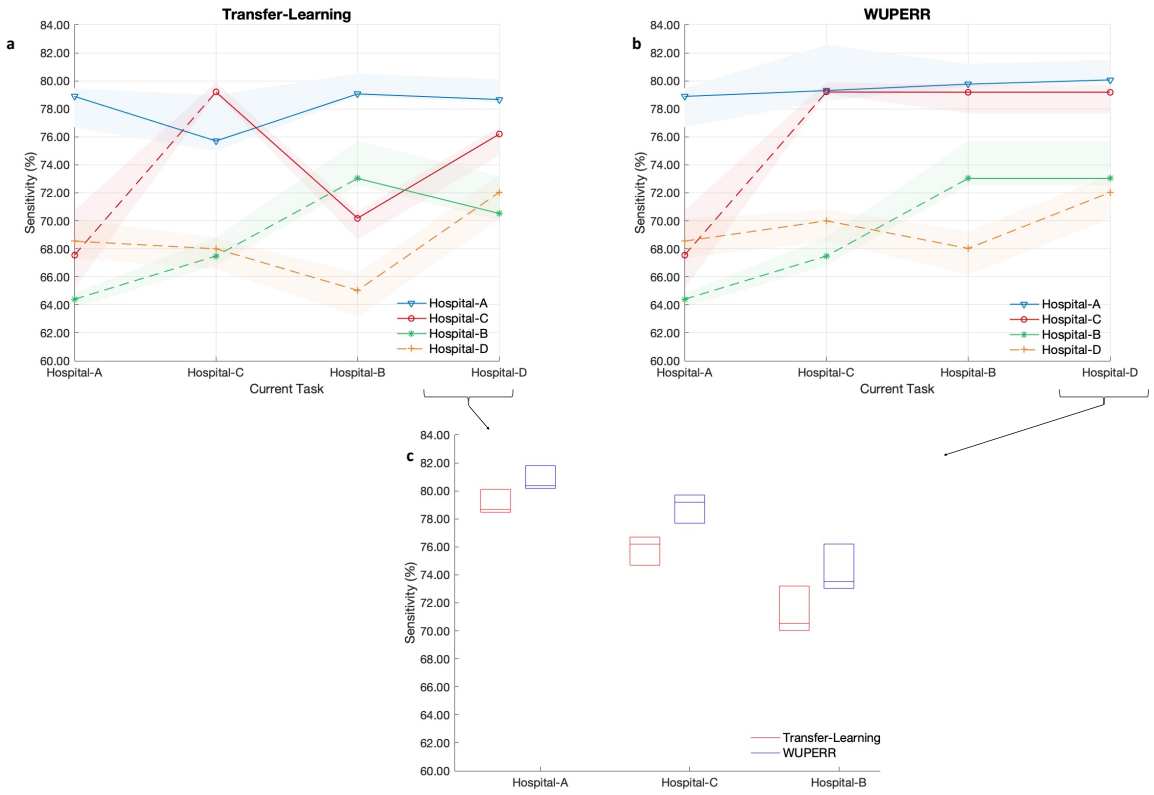

Figure 2.9. Evaluation of continual learning models for early predicting of onset of Sepsis. The performance measured using sensitivity while tasks reordered as Hospital-A, Hospital-C, Hospital-B, and Hospital-D.

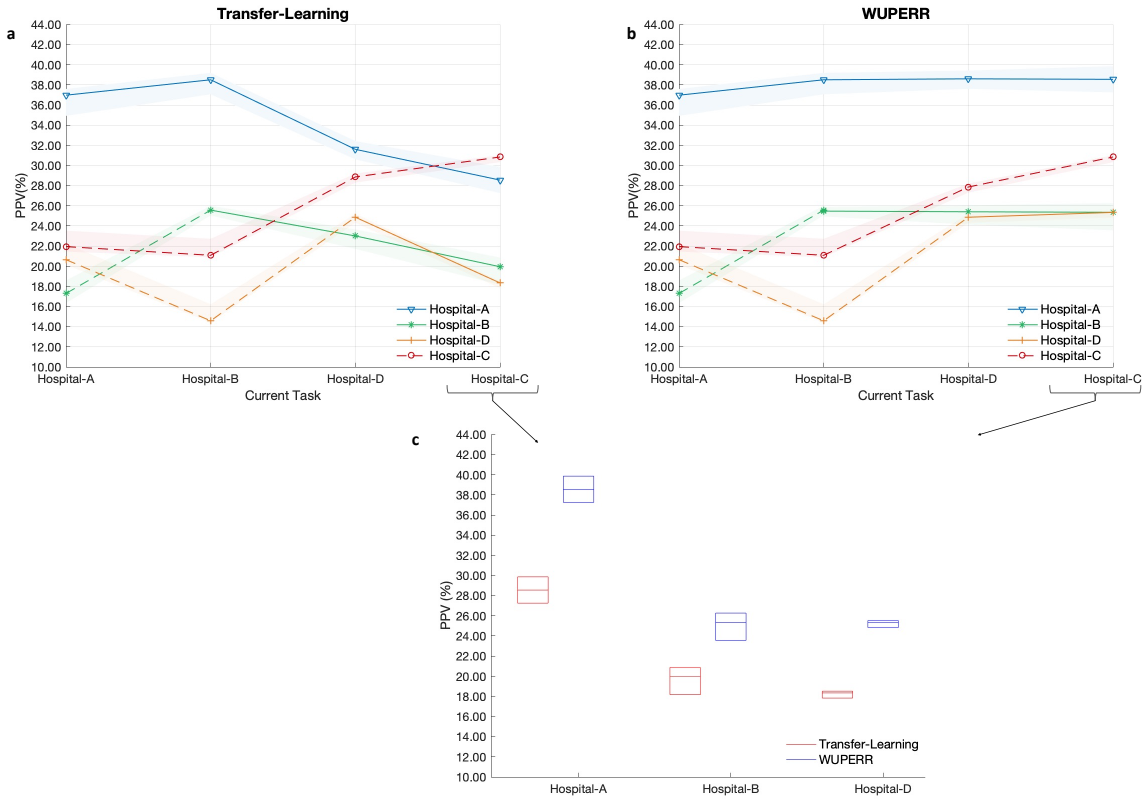

Figure 2.10. Evaluation of continual learning models for early predicting of onset of Sepsis. The performance measured using PPV metric while tasks reordered as Hospital-A, Hospital-B, Hospital-D, and Hospital-C

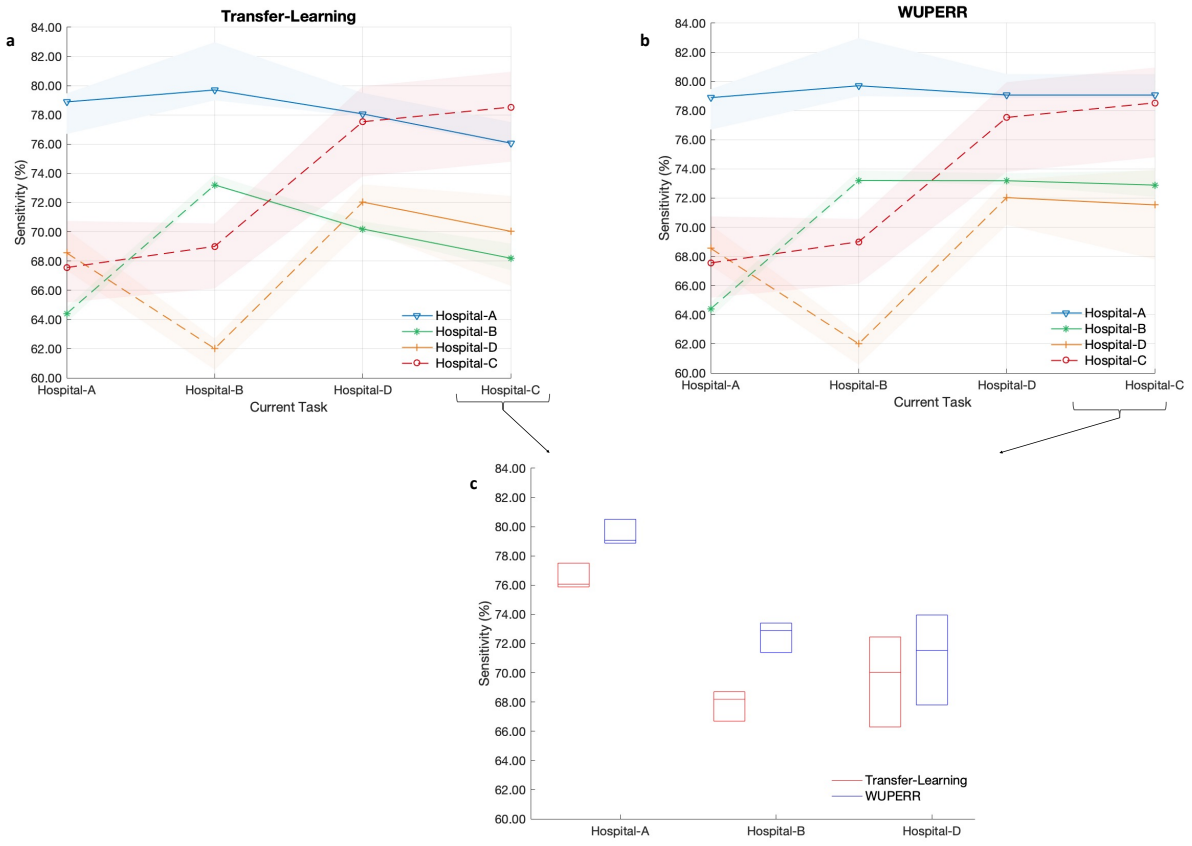

Figure 2.11. Evaluation of continual learning models for early predicting of onset of Sepsis. The performance measured using Sensitivity metric while tasks reordered as Hospital-A, Hospital-B, Hospital-D, and Hospital-C

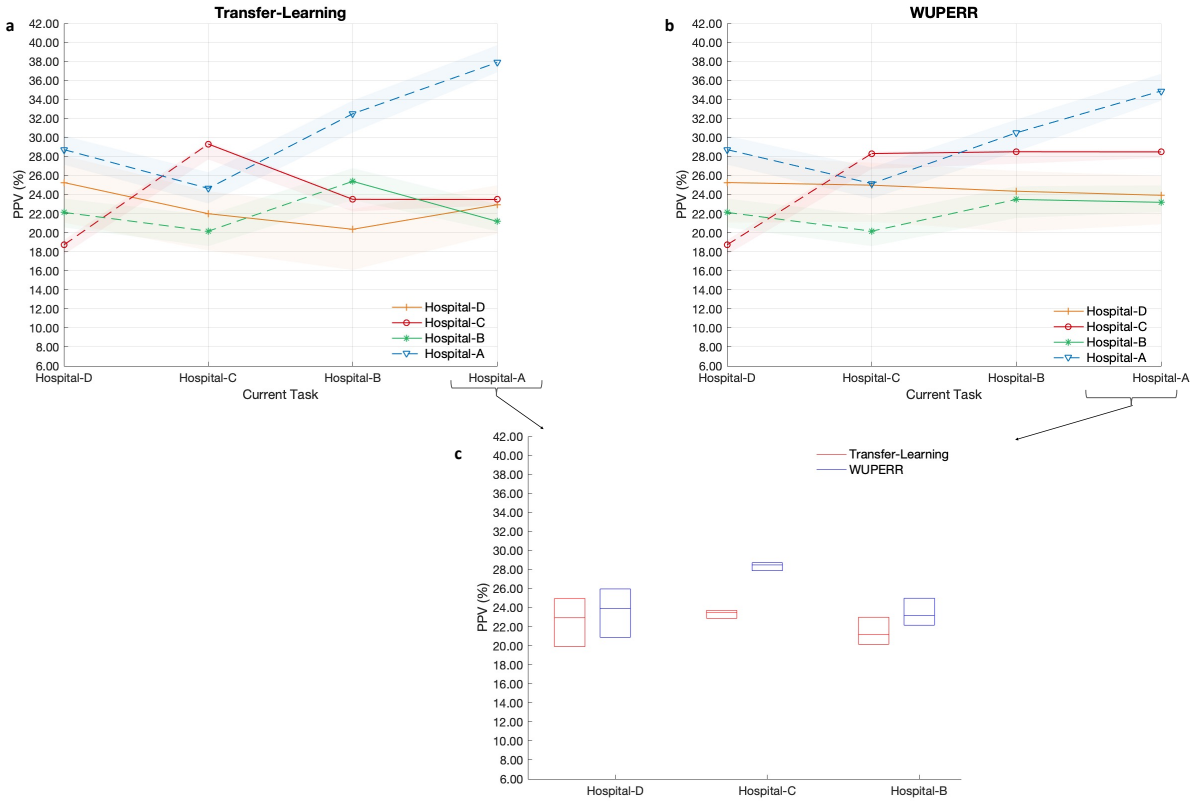

Figure 2.12. Evaluation of continual learning models for early predicting of onset of Sepsis measured using PPV metric (tasks reordered as Hospital-D, Hospital-C, Hospital-B, and Hospital-A).

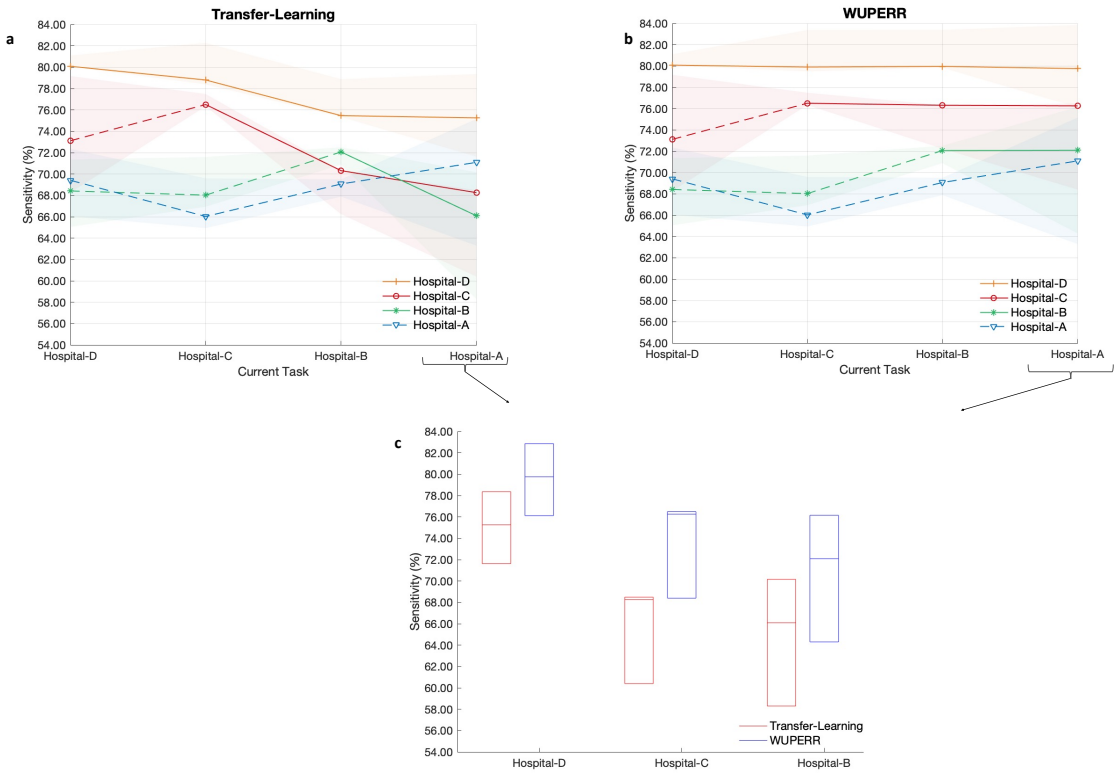

Figure 2.13. Evaluation of continual learning models for early predicting of onset of Sepsis measured using Sensitivity metric (tasks reordered as Hospital-D, Hospital-C, Hospital-B, and Hospital-A).

2.4 Discussion

In this study we designed and validated a continual learning algorithm for training generalizable clinical predictive analytics models across multiple patient cohorts. WUPERR integrates rehearsal memory with weight uncertainty propagation, and enables clinical deep learning models to learn new tasks while maintaining acceptable performance across prior tasks. We evaluated our proposed algorithm on four consecutive tasks involving early prediction of sepsis in hospitalized patients. Our results indicate that WUPERR can successfully deal with data distribution shifts that often adversely affect the generalizability of clinical predictive models. By the virtue of using data representations for continual learning, WUPERR allows the raw training data to remain at each site and therefore maintains privacy and autonomy of healthcare data. We compared WUPERR against several baselines, including Transfer Learning [51], EWC [30], and Experience Replay using three clinically relevant performance metrics, namely AUCroc, Positive Predictive Value, and Sensitivity. One may expect that learning a site-specific model should achieve the best performance, although such a model may not generalize well to external sites. WUPERR outperformed baseline Transfer Learning and EWC in terms of all three metrics to alleviate forgetting. One of the main advantages of WUPERR is the ability to learn from embedded representation of data points which makes WUPERR an appropriate approach for privacy-preserving continual learning.

Research on machine learning and deep learning has produced promising results in identification, diagnosis, and delivery of treatments in healthcare [52, 53]. Improved performance of deep learning algorithms comes at the cost of requiring large and diverse datasets [54]. However, patient privacy and data governance considerations have contributed to data silos and have made the task of constructing large multicenter datasets impractical. Some of the challenges of learning complex models from data silos have been addressed by Federated learning, where a decentralized learning algorithm relies on local model updates to con-

struct a global model [55–57]. Huang et al., introduced the community based federated learning (CBFL) framework to predict prolonged ICU stay and mortality [58]. Qayyum et al., used clustered federated learning (CFL) for identifying patients with Covid-19 [59]. In the subsequent chapter, our focus shifts to the exploration of federated learning, where we introduce an innovative aggregation algorithm tailored for this paradigm. Although federated learning is promising, but federated learning models tend to learn an average model that may perform suboptimally within any given local site. In particular, standard federated learning methods do not address the problem of data distribution shift and model drift that result from differences in patient demographics and workflow-related practices.

On the other hand, continual learning methods (such as WUPPER) allow models to incrementally learn new tasks while preserving their performance on prior tasks. This allows a model to adapt to dynamic changes and shifts in data distribution across different healthcare sites. A recent longitudinal analysis of a sepsis alert algorithm across four geographically diverse health systems reported significant dataset shift due to a change in the case-mix over time [60]. As such, algorithm monitoring [32] and continual learning are needed to ensure such systems adapt to the underlying changes in data distribution and can maintain a high level of accuracy.

This study has several limitations. The proposed learning method allows a model to adapt to shifting data distributions across clinical sites, however, a key requirement is the quality of input data and labels. Recently, conformal prediction was introduced to provide a probabilistic framework for assessing out-of-distribution samples and to detect outliers and noisy data [32]. WUPERR can be used in association with conformal prediction to control the quality of input data at each site for continual learning. In addition, differences in quality of labels at various sites can pose a challenge to continual learning. Combining WUPERR with methods for assessing and correcting label noise may provide a mechanism for training high-quality models. Moreover, WUPERR does not address the problem of

partial data availability, but recent work in continually growing neural networks can be combined with WUPERR to design algorithms that can leverage additional variables and features in new datasets [61,62].Finally, the datasets used in this study were collected from major academic medical centers and may not be representative of smaller community and rural hospitals. However, our proposed framework is likely to benefit smaller hospitals that may not have the necessary resources to maintain large clinical data warehouses, since fine-tuned pre-trained neural networks have been shown to outperform neural networks trained from scratch on smaller datasets [63]. In summary, our findings provide significant clinical evidence for the applicability of continual learning to design and update of generalizable clinical predictive models.

2.5 Acknowledgements

Chapter 2, is a reprint of the material as it appears in Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. Sci Rep 12, 8380 (2022).Amrollahi, F., Shashikumar, S.P., Holder, A.L.,Nemati,S. The dissertation author was the primary investigator and first author of this paper.

Chapter 3

Improving Clinical Deep Learning Model Generalizability via Federated Learning

3.1 Introduction

In previous chapter, we discussed the need for generalizable models, and the difficulty of curating multi-center data to train generalizable models. To protect patient privacy, it has been suggested that instead of moving data across institutional boundaries, models can be moved around and learning can occur in a decentralized manner. Federated learning (FL) is a paradigm that addresses learning with fragmented sensitive data [56, 64].

FL has emerged as a transformative approach in the realm of digital healthcare, offering a collaborative model of learning that preserve patient privacy and data security. FL circumvents the traditional barriers to data-sharing, imposed by strict privacy regulations and ethical considerations, through enabling healthcare institutions to contribute to the development of robust predictive models without the need for sharing their sensitive patient data [65].

FL through transferring the model characteristics addresses the problem of data governance and preserving data privacy for training DL models on multi-center EHR data [66]. FL

is the problem of training a shared consensus model from decentralized data. This is particularly beneficial in healthcare settings, where data cannot be aggregated centrally due to privacy concerns.

The utility of FL in healthcare is further amplified by its capacity to harness the rich, diverse datasets that are naturally distributed across different institutions. These varied datasets include a range of patient demographics, different disease prevalence, various treatment protocols, and treatment outcomes, providing a comprehensive canvas for developing generalizable models. By combining insights from various sources, FL can result in more reliable CDS models for clinical decision-making [55].

Federated learning (FL) is being explored extensively in the healthcare domain for its ability to utilize distributed datasets while maintaining the privacy of the data. [67] provided an overview of how FL can be applied to healthcare informatics for various disease prediction and classification tasks. For example, [68] provide a hypothetical study for using FL to classify breast density from mammograms in real-work setting. In [69] they use FL for brain tumour segmentation on the BraTS dataset. Cho et al. [70] proposed FL framework allows each participating site to compute summary statistics from its genetic data cohort and then shares only these summary statistics with a central server, rather than raw genetic data. In this study authors conduct Genome-Wide Association Studies (GWAS) across multiple biomedical datasets without centralizing individual-level genetic data.

FL also opens avenues for continuous learning and model refinement. As healthcare is a rapidly evolving field, the ability to continuously update models with new data from various institutions, positions FL as an ideal paradigm for digital healthcare innovation [71].

FL works in a collaborative manner, such that data remains at each health site, and a globally aggregated model would be exchanged and transferred across sites. FL models are being designed to tackle statistical heterogeneity, ensuring that the global model remains robust and performs well across all participating entities [72]. Although, a significant

concern is the non-IID (independently and identically distributed) nature of EHR data across different institutions, which can introduce bias and affect the performance of the global model [73].

Federated learning encompasses diverse architectures, each defining how the collaborative learning process unfolds among participating entities, such as healthcare institutions, and potentially a central coordinating server. The chosen architecture is pivotal, as it directly influences the efficacy, confidentiality, and operational efficiency of the learning system. The primary types of federated learning structures:

- **Centralized Federated Learning (FL):** In this structure, there's a central server that coordinates the learning process. Each Client (here a healthcare institution) computes model updates using their local data and sends these updates to the central server. The central server aggregates these updates to improve the global model and then distributes the updated global model back to the clients [74].
- **Decentralized Federated Learning (FL):** This structure removes the need for a central server. Clients communicate directly with each other to share model updates [75].
- **Hierarchical Federated Learning (FL):** This introduces intermediate aggregation nodes, which can be beneficial for large-scale applications. It typically involves a layered model where local aggregations are performed in subgroups (e.g., within a hospital network with same standard), and a higher-level aggregation might occur at the central server (e.g., across different healthcare systems) [76].

In our research, we have formulated a novel algorithm for updating the global model within federated learning systems. Our empirical results indicate that this new algorithm surpasses standard federated learning approaches, which rely on averaging model parameters. For the sake of simplicity and clarity in our comparative analysis, we adopted a centralized

architecture. Furthermore, we compare the performance of our federated learning model in predicting the rsik of sepsis against the continual learning models delineated in the preceding chapter.

Figure 3.1 illustrates the schematic diagram of the baseline FL approach. Within the FL model, the server initiates the global network's parameters randomly and shares the global weights (i.e., WG) with all the clients (i.e., Hospital A-D). At each iteration of server-client communication until convergence the following steps occur: each client fine-tunes the network on the locally kept data and communicates the new weights with the server. The server updates the global weights using the pre-defined consensus algorithms and shares the new global weights with all clients.

3.1.1 Research design and methods

3.1.2 Study Populations

For this analysis, we focused on the same problem as the prior study on predicting the onset of sepsis using four distinct datasets. To develop and evaluate the FL framework, we randomly split the train data at each center with the same rate of sepsis prevalence into five non-overlapping batches. Through the FL framework at each iteration, our clients (i.e., Hospital A-D) use one of the batches to fine-tune the local model.

3.1.3 Methods

Let k be the number of clients (here k=4, Hospitals A-D)) where the data reside. The Server randomly initialized our neural network and shared the parameters with all clients. At each iteration each client fine-tunes the network locally and they communicate the local updates with the server.

Let D_k denote the data distribution associated with client k (for a total of K nodes) and n_k the number of samples available from the client k. Then $N = \Sigma_{k=1}^{K}(n_k)$ is the total sample

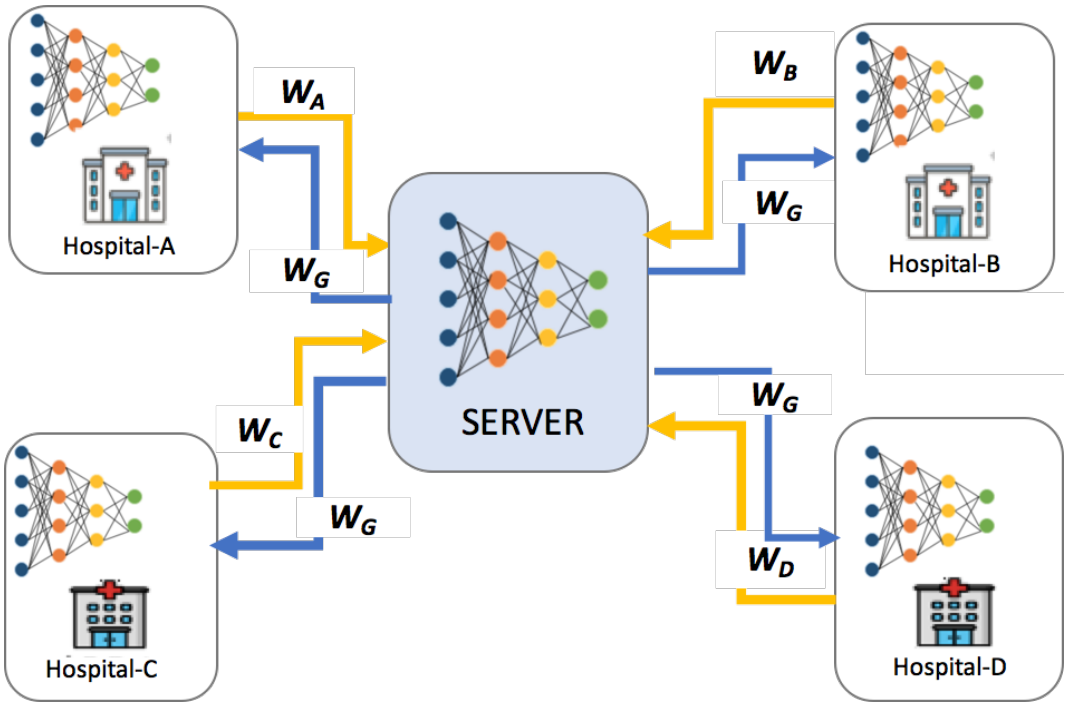

Figure 3.1. Block diagram of the Centralized Federated Learning (FL) approach. At each iteration all the clients (i.e Hospitals A-D) send the locally fine tuned parameters to the server. Server further shares the sum of weighted scaled parameters with all the clients.

size and l denotes global loss function obtained via a weighted combination of K local losses. The FL aims to minimize the l (see equation.3.1):

$$min_{W_G} l(W_G) = \Sigma_{k=1}^{K} l_k(W_G) \tag{3.1}$$

– **Vanilla Federated Learning (baseline):** In this approach the global model parameters (W_G) is the weighted average of fine-tuned model parameters at each client w_k.

$$W_G = \Sigma_{k=1}^{K} (\frac{n_k}{N}.w_k) \tag{3.2}$$

– **Novel Federated Learning Approach:** In this approach, after each iteration t all clients send their fine-tuned model parameters w_k^t to the server. The server then update global weight W_G^t according to scaled sum of w_k^t.

$$W_G^t = \Sigma_{k=1}^{K} (S_k^t.w_k^t) \tag{3.3}$$

S_k^t the scaling factor for weights of client k at iteration t, calculated based on model performances on external test data at server. The external test data on servers comprises unseen data from all clients. We calculate S_k^t as:

$$S_k^t = \frac{e^{AUC_k}}{\Sigma_{k=1}^{K} e^{AUC_k}} \tag{3.4}$$

AUC_k is the average performance of the k-th model on test data from all clients. As such, this approach rewards local weights that have a higher overall generalizability.

3.2 Results

We evaluated and compared the performance of our new FL approach with the vanilla FL algorithm using the C-AUC metric. Figure 3.2 shows the performance of the global network updated using the two consensus algorithms on test data after each iteration. Test data comprises unseen data from all clients and we show the model performance for test data from each client in a different color and marks. We observed that our novel approach outperformed the vanilla FL for Hospital-A, Hospital-B, and Hospital-D. We hypothesize that our novel approach for updating the global network could not perform well on Hospital-C because of the drift in Hospital-C's underlying data distribution. Fine-tuned network parameters from Hospital-C could not perform well on all other hospital's unseen data so our approach did not integrate parameters from Hospital-C while the vanilla FL model that averages the weights integrate more parameters from Hospital-C, and consequently could not perform well on all other hospitals. So our approach is more robust against the adversarial data from one client.

We further compare the FL algorithms with WUPERR on the same test data from each hospital after the model trained on all four hospitals' data. Figure 3.3 indicates WUPERR outperformed both FL algorithms on all four hospitals after the model trained on data from all hospitals, mainly because WUPERR improves the model's generalizability through broadening the model's knowledge while transferring and maintaining the previously acquired knowledge. On the other hand, the FL approaches are prone to forgetting previously learned tasks.

3.3 Discussion

The incorporation of Federated Learning (FL) in healthcare domain has been a burgeoning area of interest, yielding promising results in protecting patient privacy while enabling collaborative learning across different institutions. Prior works in this domain have highlighted FL's potential in creating robust predictive models for various clinical outcomes without the need for data centralization. Sheller et al. [77] demonstrated FL's utility in brain tumor segmentation tasks, showcasing the model's adeptness at learning from distributed data sources while maintaining data confidentiality. Similarly, in [78] Brisimi et al. highlighted the application of FL in patient-specific heart rate prediction models, indicating that FL can cater to individual patient characteristics while leveraging a broad dataset.

Our work extends these foundational studies by introducing a novel algorithm for updating the global model in the context of sepsis prediction. When contrasted with traditional FL approaches that rely on model parameter averaging, our approach demonstrates an enhanced ability to cope with the intrinsic variability of healthcare data across different hospitals. Furthermore, our method demonstrated superior performance in predicting sepsis in hospitals A, B, and D. This suggests that the adaptive mechanism of our algorithm for updating the global network is more effective, especially in environments where data distributions can significantly diverge, as in the case of Hospital-C.

In the FL landscape, especially within healthcare, our study contributes to the evolving narrative of how models can be adapted to address the challenges of data heterogeneity and drift, which are especially pronounced in multi-institutional healthcare datasets. The results from our investigation highlight the potential of FL in creating more tailored and adaptive models that do not merely learn across settings but also resist the dilution of performance that often accompanies the incorporation of outlier data. This exploration

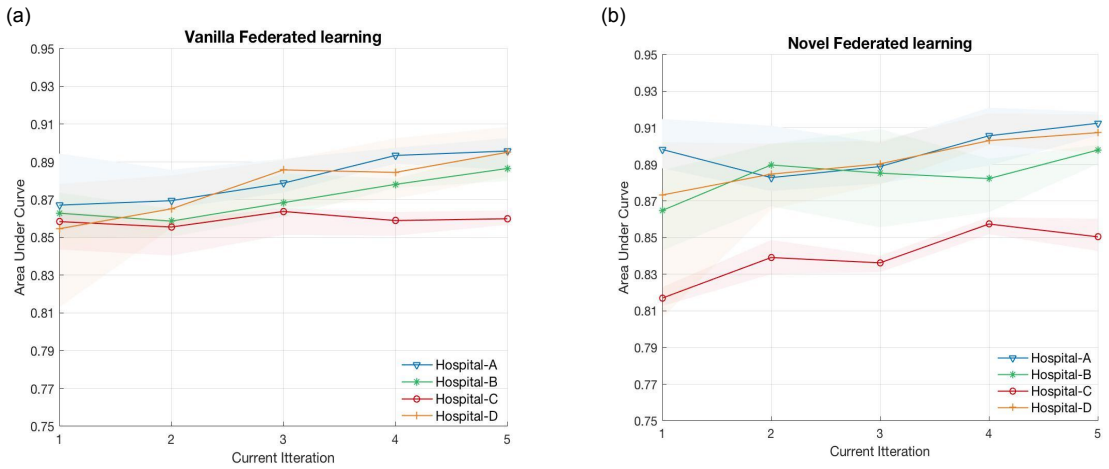

Figure 3.2. Evaluation of Federated Learning (FL) models for predicting onset of sepsis using AUC metrics. Panel (a) represents the model performance on test data using vanilla FL after each iteration. Panel (b) illustrates the model performance on test data using our novel approach for updating the global weights after each iteration.

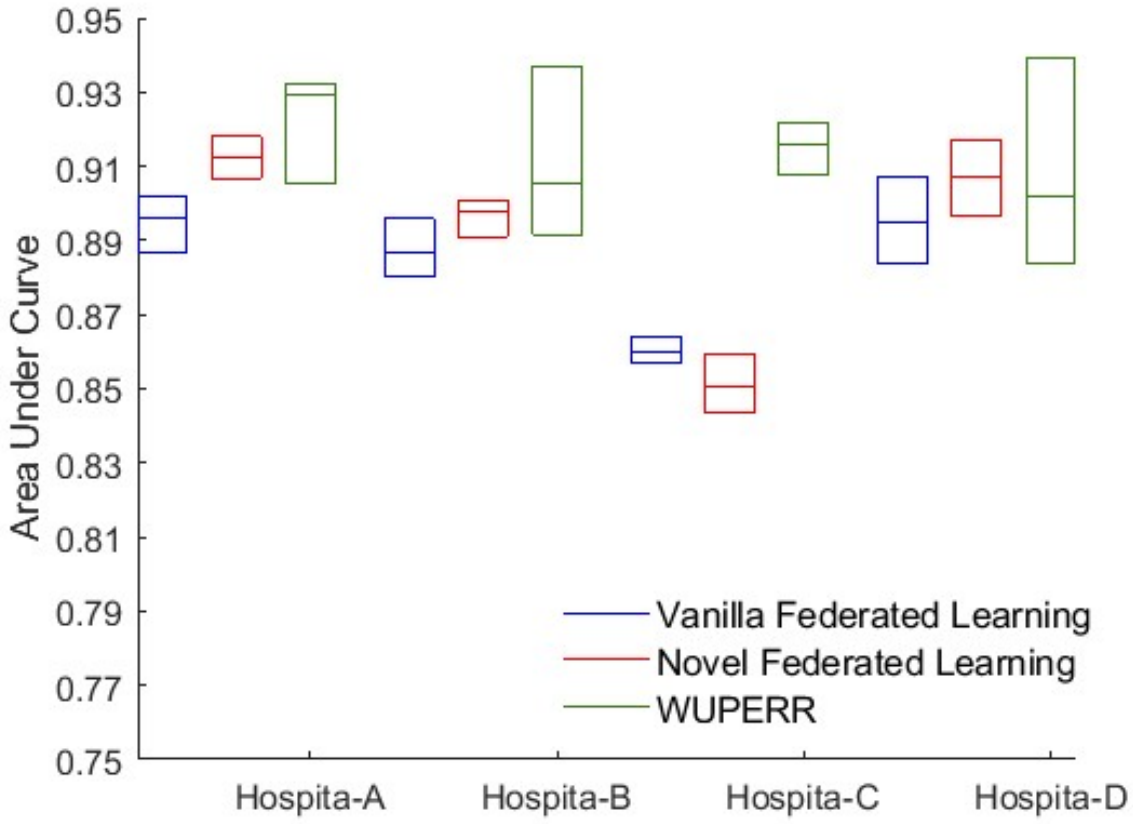

Figure 3.3. Comparing the model performance on predicting onset of sepsis on test data of four hospitals after model learned data from all hospitals using proposed WUPERR algorithm, new proposed Federated learning model and vanilla federated learning.

into FL's adaptability and resilience paves the way for further progress in the predictive analytic technology, ensuring that it continues to align with the complex and dynamic nature of healthcare data.

Furthermore, when compared with the WUPERR algorithm, it is apparent that while our FL method advances the capabilities of FL, WUPERR provides an even more enhanced performance after it has sequentially assimilated data across the four hospitals. WUPERR's strength lies in its ability to acquire new information without forgetting previously acquired knowledge, a feature that is paramount in healthcare settings where continuity of information is crucial. This continuous learning approach aligns well with the dynamic nature of healthcare data, where each patient or hospital may introduce new patterns to the model.

Chapter 4

Generating synthetic longitudinal electronic health records for machine learning applications

4.1 Introduction

The advent of Generative Adversarial Networks (GANs) in 2014, as introduced by Goodfellow et al. [79], was a major turning point in the field of machine learning. GANs comprise two distinct neural networks that engage to improve their functions: the generator, which creates data, and the discriminator, which distinguish between real and fake data. This unique architecture allows GANs to generate new data instances that are similar to real instances, thus achieving impressive results in generative tasks.

The generator network are able to produce synthetic data that the discriminator network cannot distinguish it from real data. Meanwhile, the discriminator learns to differentiate between the generator's fake data and the real data it has been trained on. Through iterative training, both networks incrementally improve their performance; the generator progressively creates more realistic data, while the discriminator enhances its ability to identify subtle differences that distinguish real data from generated one.

The original GAN framework introduced in [79] has undergone various enhancements to

improve stability and output quality. These include modifications such as the introduction of Deep Convolutional GANs (DCGANs) by Radford et al. [80], which incorporate convolutional neural networks into the GAN architecture, improving the quality of generated images. Other notable innovations, such as the Wasserstein GAN (WGAN) [81], address training stability issues by modifying the loss function used to train GANs, thereby providing more reliable convergence during training.

The training process of GANs involves a finely tuned balance between its components. If the discriminator becomes too proficient, the generator may fail to improve due to discouraging gradients. Conversely, a too-powerful generator may produce data that doesn't incorporate the diversity and variability seen in real data, a phenomenon known as mode collapse [82]. Advances in GAN training methods, such as the introduction of auxiliary classifier GANs (AC-GANs) [83], have aimed to mitigate these issues by incorporating additional information, such as class labels, to guide the generative process more effectively.

GANs have achieved remarkable results in producing synthetic data sets in industry. These models are designed to synthesis data from an initial random noise vector or based on specific features to which the model is conditioned. With the aim of maintaining privacy, the foundational principle here is that the synthetic data, generated from random inputs, must be indistinguishable from actual data. Within the realm of generative models, GANs have become notably prominent due to their capacity to generate highly persuasive samples that closely resemble the distribution of actual data. GANs have been successful in generating complex, high-dimensional data, with proven effectiveness across domains such as imagery [84], audio [85], text [86], and sequential time-series [87].

4.1.1 GANs in Healthcare

The increased adoption of EHR has created unprecedented opportunities for advancing healthcare analytics. However, the sensitive nature of medical data poses significant privacy challenges that restrict the free exchange of data for research purposes.

In healthcare, the sensitivity and privacy of patient data necessitate solutions that can utilize data without compromising patient privacy. GANs have emerged as a powerful tool to address this issue by generating synthetic, yet realistic, EHR data that can be used for research without compromising patient confidentiality. GANs enable researchers to overcome privacy constraints while advancing medical research and training AI models. Choi et al. [88] demonstrated the utility of GANs in creating synthetic patient records, paving the way for a new era of privacy-preserving data sharing.

Synthetic EHR data generation with GANs not only aids in protecting patient privacy but also enhances the ability to model complex and rare diseases by creating large, diverse, and balanced datasets that may not be available in real-world settings due to rarity or under-reporting. Esteban et al. [89] showcased the successful application of GANs to create continuous EHRs, which enabled the development of predictive models for clinical events. The synthetic data generated by these models can accurately reflects the complex time-related patterns observed in actual patient data, helping in the creation of dynamic patient profiles for various simulation and modeling tasks.

Another challenge in medical data analysis is the imbalanced nature of datasets, particularly when it comes to rare diseases or outcomes. GANs can help overcome these limitations by generating additional synthetic examples of underrepresented classes, leading to improved predictive modeling. The work by Jones et al. [90] exemplifies how synthetic data can be used to augment datasets, creating more balanced data that enhance the performance of classification algorithms. This methodological advancement is particularly relevant for

training robust machine learning models that can better predict rare clinical events.

Furthermore, the potential of GANs to generate dynamic EHR data is especially valuable in addressing the temporal and sequential nature of healthcare data, where patient states can evolve over time. This ability of GANs to generate dynamic EHR data is transforming personalized healthcare. By leveraging temporal and sequential data synthesis, GANs can model the progression of diseases and patient responses to treatments over time. This is critical for personalized medicine, which focuses on customizing treatments to suit the unique disease progression of each individual patient. For instance, the Recurrent GAN (RGAN) and Recurrent Conditional GAN (RCGAN) proposed by Esteban et al. [89] are specifically designed to generate sequential data that maintain the integrity of temporal correlations, thereby facilitating more accurate predictive analytics for patient trajectories and treatment outcomes.

The Temporal GAN (TGAN) proposed by Yoon et al. [87] illustrates how EHR data can be synthesized to retain temporal relationships, allowing for more accurate modeling of patient states over time. This is not only beneficial for predictive modeling but also it help researchers for simulating patient responses to various treatment regimens, aiding them in decision-making and treatment planning.

In the domain of synthetic EHR data creation, the use of GANs comes with its own set of intricate challenges. In our study, we use GANs to fabricate synthetic EHR data, ensuring that the privacy of sensitive information is upheld. Through this process, we maintain the balance between data utility and patient data confidentiality, addressing the nuances that arise when GANs are applied to health-related data, which is inherently complex and multifaceted.

The diagram depicted in Figure 4.1 illustrates the architecture of our proposed model, which is meticulously engineered to generate high-fidelity synthetic EHR data. This model, inspired by a framework presented in [91], is designed to retain the critical statistical

properties that are essential for subsequent analytical tasks while ensuring the stringent privacy requirements of the original dataset are met.

Our model integrates a robust encoding and decoding mechanism that capably handles the features characteristic of EHR data. It adeptly normalizes complex and skewed distributions, accurately representing even the missing data points, thereby maintaining the integrity and utility of the synthesized dataset. To evaluate the efficacy of our synthetic data generation method, we employed the comprehensive and diverse AllofUs dataset, spanning multiple institutions. We then compared the performance of models using this synthetic data in predicting the onset of sepsis against those using the original dataset.

Digging further, we have customized this model to enhance the development of generalized predictive models in a clinical setting. It is particularly effective in incremental learning, that involve sequential streaming data, a challenge we thoroughly investigate in previous chapters. By leveraging synthetic EHR data, our model not only circumvents privacy concerns but also contributes to the continuous improvement of healthcare predictive analytics. This dual advantage showcases the potential of our model to serve as a pivotal tool in the advancement of digital healthcare, propelling forward the capabilities of clinical decision-making aids.

4.2 Methods

4.2.1 Study Populations

In our study, we utilized patient data from the AllofUs Research Program, supplied by the National Institutes of Health (C2022Q4R9). We secured the required approval from the Institutional Review Board (IRB) under the AllofUs program (Protocol Number: 2016-05, approval date: March 17, 2021). An exemption was also granted by our institution's IRB, allowing us to proceed without the need for individual consent, provided that all research

was conducted in line with the ethical standards of the Declaration of Helsinki, and under the oversight of the relevant human research ethics committees.

The dataset is a multicenter cohort that spans 35 US hospitals, covering over 331,382 individual patient records. The AllofUs program aims to encompass a broad cross-section of the US populace, including minority groups often not sufficiently represented in biomedical research.

The AllofUs dataset's rich diversity, encompassing a wide array of patient demographics and information from an extensive network of healthcare institutions, renders it an exceptional resource for our research. This is a key factor in our choice of the AllofUs dataset for our study, as it aligns with our objective to create universally adaptable healthcare models.

The dataset in question comprises a comprehensive collection of EHR data, such as patients' medical histories, vital sign measurements, and laboratory test outcomes. Additionally, it includes demographic information such as gender, race, and ethnicity, as reported by the participants themselves. A more in-depth exploration of the AllofUs dataset can be found in the specialized report produced by the AllofUs Research Program [92], which delves into the nuances and structure of the collected data.

4.2.2 Data Preprocessing

To be consistent with our prior study cohort, similarly we included all patients 18 years or older who developed sepsis as defined by the Third International Consensus Definition of Sepsis ("Sepsis 3") during hospitalization [9]. We focused on sequential hourly prediction of sepsis starting at hour four after admission. Patients who were identified as having sepsis prior to prediction start time or those with no measurement of heart rate or blood pressure prior to the prediction start time or those whose length of stay were more than 21 days were excluded. The same clinical variables as 2.1 and 50 histories of medication were

extracted. All variables were obtained via automated Observational Medical Outcomes Partnership (OMOP) common data model queries provided by Amrollahi et al [93]. All clinical variables were extracted from the AllofUs database using OMOP concepts codes. For variables with multiple concept codes with distinct measurement units we applied appropriate conversion rates.

All vital signs and laboratory variables were organized into 1-h and 1-day non-overlapping time series bins to accommodate for different sampling frequencies of available data for the sepsis cohort. All the variables with sampling frequencies higher than once every hour (or day) were uniformly resampled into 1-h (or 1-day) time bins, by taking the median values if multiple measurements were available. Variables were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). Mean imputation was used to replace all remaining missing values (mainly at the start of each record).

In the preceding chapter, we enhanced model performance by incorprating the slope of local trends and TSLM for each vital sign and laboratory measurements. In this chapter, to generate synthetic EHR data, we account for all variables and the patterns of missing data on an hourly basis. We also used this missingness information to calculate the rate of change and TSLM on generated data for downstream task.

The patient characteristics of the cohorts have been tabulated in Table 4.1. All continuous variables are reported as medians with 25% and 75% interquartile ranges (IQRs). Binary variables are reported as percentages.

4.2.3 Stochastic Normalization and Feature representations

Our EHR data consist of 191 total features including both static and longitudinal variables. Each static and temporal feature can be categorized into either numeric or categorical.

Table 4.1. Summary of patient characteristics of the AllofUs dataset

	Non-Septic	Septic
Patients (#)	25304	518
Male (%)	7692 (30%)	193 (37%)
Median Age [IQR]	48.0[33.0-58.0]	51.0[41.0-59.0]
Race (%)		
Caucasian	12627(50%)	272(53%)
African American	5966 (24%)	94 (18%)
Asian	810(3%)	8(2%)
TSepsis [IQR]	-	11.0 [8.0 -29.0]

we have the four categories of features for each of the patient including: measurement date and time, static numeric feature Sn (e.g., age), static categorical feature Sc (e.g., medications), and time-varying numerical feature (e.g., vital signs). Note that each patient record may have a different sequence length. We set the maximum length of sequence as 36. For longer sequences, we only use the last 36 time steps. One-hot encoding is used for categorical features.

With all these features, given training dataset (D), EHR records for patient (i) can be represented as:

$$D = \{S^n(i), S^c(i), M_\eta(i), V^c_\eta(i), V^n_\eta(i)\}_{\eta=1}^{\eta=T(i)} \tag{4.1}$$

Where N is the total number of patients and T is the sequence length of EHR records for patient i. At each time stamp η, V^c represents the time varying categorical features, and V^n is the time varying numerical features. M contains missing patterns at each time-stamp for longitudinal variables.

EHR data records contain feature distributions where the mass probability is condensed within a small numerical range, this is a severe issue for training the Generative model (known as mode collapsing [94]). GAN models suffer from mode collapse and would

have the tendency to generate common values for all samples. To circumvent GANs' overemphasizing on the generation of some commonly observed data values we adopted stochastic normalization technique used by yoon et al [91].

The normalization method maps the raw feature distributions to more uniform distribution that is easier to model with GANs. This method is reversible. Stochastic normalization maps the original feature space into a normalized feature space (with uniform distribution), and then the applied renormalization recreates the skewed distributions. We have found that applying Stochastic normalization to our dataset improves the model performance. Figure 4.2 illustrated a pseudocode of stochastic normalization and renormalization algorithms.

To evaluate the framework we divide the patients into disjoint train and test datasets with 90% and 10% ratios. We only use the training split to train our framework to generate synthetic data. At inference, we compared the model performance on test data while trained on generated synthetic data and real train data.

4.2.4 Encoder-Decoder Architecture

We used an encoder-decoder to jointly extract the representations from multiple types of data, including static, temporal, measurement time, and mask features. We encode these heterogeneous features into joint representations from which the synthetic data samples are generated. An encoder-decoder model is beneficial for GANs model convergence as it condenses high-dimensional heterogeneous features into latent compact representations. The encoder model uses the static, temporal, and missing pattern of data and generates the encoder states (E). The decoder would further decode the encoder states (E) to generate normalized encoder inputs. Teacher forcing during training where the target (input to the encoder) is passed as the next input to the decoder is employed to help the model converge faster. Zero padding and masking technique has been used to capture the dynamic of sequence length. If the decoder model can recover the original heterogeneous

data correctly, it can be inferred that E contains most of the information in the original heterogeneous data.

4.2.5 GAN Architecture

The trained encoder model is further used to map raw data into encoded representations, that are then used for GAN training. So the GAN learns to generate encoded representations that can be decoded into raw data. We first utilize the trained encoder to generate encoder states (E) using the original normalized raw data. Next, we use the GAN framework to generate synthetic encoder states ($\bar{E}$) to make synthetic encoder states close to the real encoder states. The generator uses the random noise vector (Z) where $Z \sim N(0,1)$ to generate synthetic encoder states. Then, the discriminator tries to distinguish the original encoder states E from the synthetic encoder states $\bar{E}$. As the GAN framework, we adopt Wasserstein GAN with Gradient Penalty due to its training stability for heterogeneous data types.

After training the both encoder-decoder and GAN, we can generate synthetic EHR data from any random vector. Note that only the trained generator and decoder are used at the inference time. The trained generator uses the random vector to generate synthetic encoder states ($\bar{E}$). Then, the trained decoder uses the synthetic encoder states as the inputs and last time predicted synthetic temporal, static and missingness EHR values to generate new time point EHR synthetic data.We have an extra post-process module to renormalize generated synthetic EHR data to raw synthetic EHR data for the downstream tasks.

4.2.6 Evaluation of GAN performance

Evaluating the performance of GANs on approximating data distribution is an important challenge and active research area in training GANs. Although there are popular methods such as Inception Score (IS) [95], Fréchet Inception Distance (FID) [96], and Perceptual

Path Length (PPL) [97] for evaluation the quality of generated images and videos by generative models, there is no well-known method to assess the quality of time-series generated data. IS and FID use the already pretrained image classification model (inception v3) and remove the last layer of model to get the representation of images. These methods further use KL-divergence of the representation for generated images and real images as quality metrics of generated methods.

To monitor the quality of the generated data with every 5 epoch we plot the distribution of the generated data versus the real encoded data using T-SNE. T-SNE is a nonlinear dimensionality reduction method which can deal with linearly non separable data. Figure 4.3 illustrates the generated data distribution and real data distribution for AllofUs non septic population cohort before and after training.

4.3 Results

As previously discussed, one of the most prominent goals for GANs is to benefit the future downstream analyses in the real clinical application. In this study, we worked on a relevant question on predicting the onset of sepsis four hours ahead for hospitalized patients. We observed that a single generative framework is not able to capture the rare abnormal trends occurring for septic patients, so we utilized a separate framework for non-septic patients and fine-tuned the copy of the framework specifically for septic patients. To assess the quality of generated synthetic data, we compare the model performance (C-AUC metric) trained on synthetic EHR data versus the model trained on real EHR data for predicting the onset of sepsis using real EHR test data. Table 4.2 tabulates both model performances for predicting sepsis.

Table 4.2. Model performance on predicting the onset of sepsis on Real EHR data using Real EHR data and Generated EHR data

	Performance on Test Set (C-AUC)	Performance on Test Set ($Specificity^*$)
Model I (Train using Real Data)	0.91	0.77
Model II (Train using Generated Data)	0.83	0.54

* Specificity at Sensitivity =0.8

4.4 Improving Clinical Deep Learning Model Generalizability using Generating Synthetic Electronic Health Records

We further integrate our proposed generative models with the context of aforementioned continual learning problem. In our continual learning approach, we replayed the representations of patients' episodes to prevent catastrophic forgetting while the model is trained on data from the new institution. Transferring representations of data from prior tasks may raise concerns of privacy preservation. In this study, we address this concern through transferring and replaying the generated synthetic data at each institution.

Building on the findings from the previous chapter, our analysis indicated that Hospital-C's data distribution was significantly varied compared to the other institutions. Due to the high computational demands, we focused our assessment on the data from Hospital-B and Hospital-C, which were representative of the sequential learning problem. This decision allowed us to maintain computational efficiency without sacrificing the essential components of our study. Within each healthcare facility, we use the same frame work as described in the preceding section for the creation of synthetic EHR data, encompassing both septic and non-septic patient groups. We found that the removing the Stochastic Normalization process from our frame work here did not adversely affect the performance of our model. This insight simplifies our approach and reduces the computational expenses

without compromising the integrity and effectiveness of our results.

To mitigate the issue of catastrophic forgetting associated with incremental learning, we have employed a strategy of fine-tuning our generative model for each healthcare facility involved in the study—specifically Hospital-B and Hospital-C. Our method involves transferring the learned model weights from one hospital to the next, and replaying synthetic data from the previous hospitals. This process helps in maintaining knowledge consistency across the learning phases. Additionally, our framework is designed with the capability to adjust the synthetic EHR data according to varying sepsis incidence rates. For the purposes of this study, we generated a balanced dataset comprising 2000 synthetic samples at each facility, with an equal sepsis rate of 50%, to ensure a uniform data structure when the model is applied to subsequent new hospital data.

Table 4.3 tabulate the performance of our framework. In this study we use 80% -20% split of data for model training and evaluation respectively. We first train our model on Hospital-B data (i.e Task 1), we further transfer the model weights and fine-tune the model on hospital-C data (i.e Task 2). Within the Generative Replay model, we employed our GAN framework within Hospital-B to generate synthetic EHR data, which was then replayed at proceeding Hospital-C. This approach allowed the predictive model to retain and leverage the knowledge acquired from previous institutions. We compare the proposed approach with the baseline transfer learning method, as outlined in the initial section. We assessed and reported the model's predictive performance for predicting the onset of sepsis at each hospital, utilizing both current and prior institution test data using the PPV metric.Table 4.3 reveals that replying the synthetic data while model trains on data from the new institution aids the model to address catastrophic forgetting.

Table 4.3. Evaluation of Synthetic Data Replay models for early predicting of onset of Sepsis measured using PPV^* metric

Transfer Learning		
	Hospital-B	Hospital-C
Hospital-B	39	34
Hospital-C	5	44
Synthetic EHR Replay		
	Hospital-B	Hospital-C
Hospital-B	39	37
Hospital-C	5	52

* PPV at Sensitivity =0.8 on Hospital-B

4.5 Discussion

Patient data privacy concerns are among the key bottlenecks for the sharing and exchanging of EHR data to develop generalizable ML-based models. AI innovations have tremendous potential in the clinical domain. Although patient data privacy is one of the obstacles in broad employment of AI in clinical applications. One possible solution for sharing data beyond institutional boundaries is to share anonymized data. Although, there is no single standardized set of recommendations on how to anonymize clinical datasets for sharing such that the disseminated data is protected against privacy attacks. The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) outlines Safe Harbor, and Expert Determination policies for protecting anonymity. The Safe Harbor policy enumerates eighteen identifiers that must be removed from data, based on the Expert Determination policy, an expert needs to certify that the shared data poses a low privacy risk.

Conventional methods to anonymize data, including perturbation via microaggregation, data swapping, or rank swapping, Suppression, Data masking and Differential Privacy, are expensive and tedious. Further, anonymization methods can distort important features from the original dataset, decreasing the utility of the data significantly, and they can be

susceptible to privacy attacks even when the de-identification process is in accordance with existing standards. A promising solution to sharing of EHR data is to use synthetic EHR data. Recent advances in generative GANs and their variants pave the way to generate synthetic data for a wide range of clinical applications. In this chapter, we developed a GAN for generating synthetic EHR data. We focused on key challenging aspects of real-time EHR data, including heterogeneity, sparsity, coexistence of numerical and categorical features with distinct characteristics, and time-varying features. We adopted a generative modeling framework, for generating highly realistic synthetic EHR data.

Although, there are several proposed models for time-series synthetic data generation including TimeGAN [87], RC-GAN [89], C-RNN-GAN [98]), but these alternative methods are not designed to handle all the challenges we addressed through our model, such as varying length sequences, missingness and joint representation of static and time-varying features. Our model is based on a two-stage approach that consists of sequential techer forcing encoder-decoder networks and generative adversarial networks to address these limitations. Within this model we employed mask modeling to deal with dynamic sequence length of EHR data.

To evaluate the effectiveness of our framework, we conducted two distinct experiments. In the initial experiment, we assessed the synthetic EHR data quality generated by our framework to ensure its efficacy for clinical decision support (CDS) tasks. Utilizing the extensive AllofUs multi-institutional cohort, we aimed to predict the onset of sepsis, demonstrating that models trained on synthetic data offer satisfactory performance on real-world EHR datasets. Subsequently, we investigated the potential for enhancing the generalizability of deep learning (DL) predictive models by transferring and replaying synthetic EHR data. Our findings indicate that incorporating synthetic EHR data allows our models to progressively learn from a continuous data stream while retaining previously acquired knowledge, thereby mitigating the problem of catastrophic forgetting.

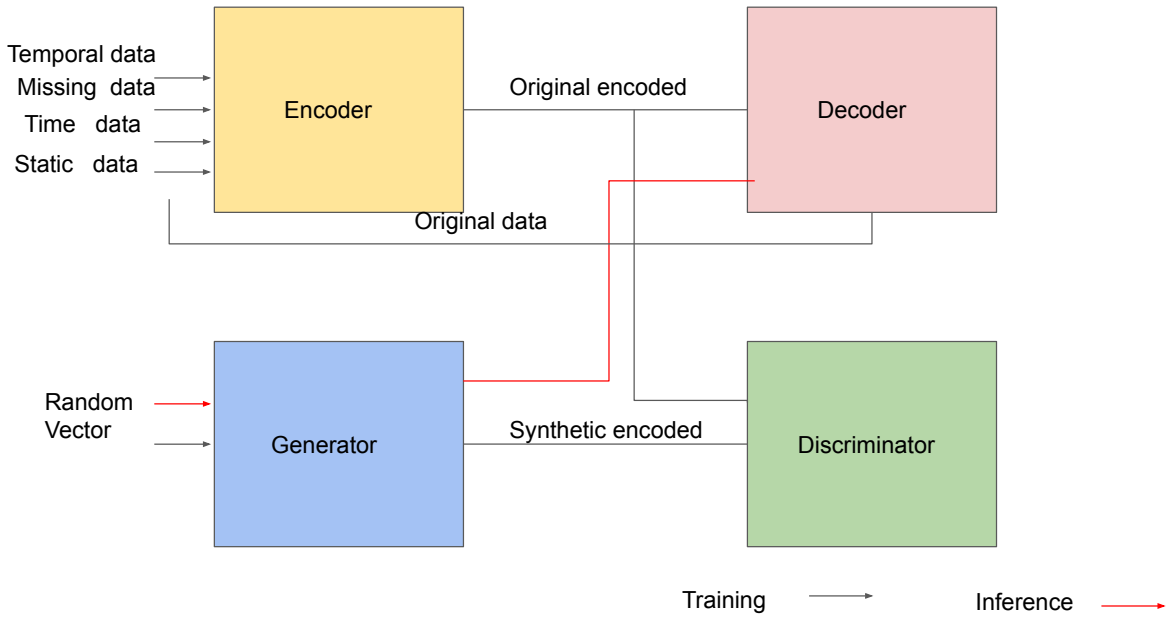

Figure 4.1. Block Diagram of generating synthetic EHR data from the original data.At inference, we only use the trained generator and decoder to generate synthetic data (shown in red arrows)

Stochastic Normalization Algorithm:

Input: Input data (X),Normalized_Feature=dict()
For each feature A in X do:
Uniq(A) = Unique values of A, N = Length of (A), lower-bound = 0.0, upper-bound = 0.0, Ā= A
For val in Uniq(A) do:
Find index of A whose value = val as idx(val)
Compute the frequency of val as ratio(val) = Length of idx(val) / N
upper-bound = lower-bound + ratio(val)
Ā[idx(val)] ~ Uniform(lower-bound, upper-bound)
params[lower-bound, upper-bound] =val
lower-bound = upper-bound
End
Normalized_Feature[A]=params
Return Normalized_Feature,Ā

stochastic renormalization:

Input: Normalized feature (Ā),
normalization parameters
(params=Normalized_Feature[Ā])
Unique(Ā)=unique values of Ā
A=Ā
For each val in Unique(Ā):
Find the keys (k) that val is in the boundary
idx(val)=index of Ā where value=val
A[idx(val)]=params[kl
Return A

Figure 4.2. pseudo code of Stochastic Normalization/Renormalization. The proposed algorithm can be highly effective in transforming features with high skewed distribution into approximately uniform distributions while allowing for perfect renormalization into the original feature space.

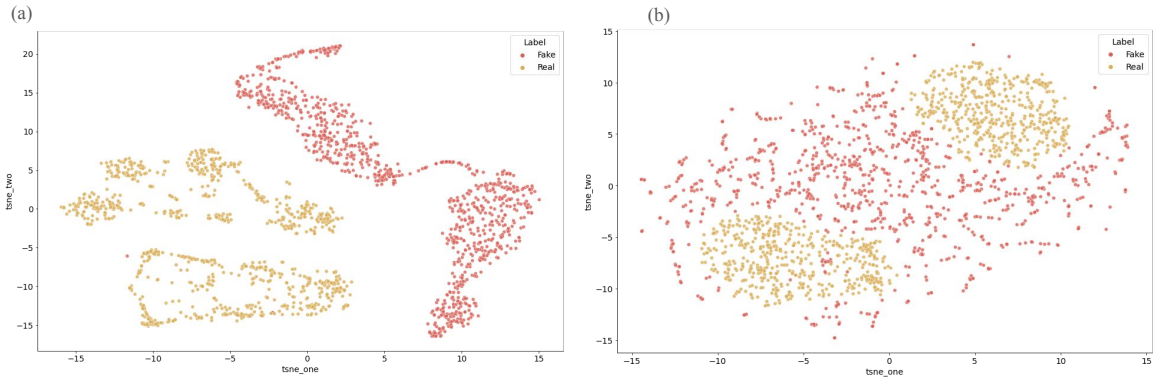

Figure 4.3. T-SNE plot of generated data versus representation of real data for non septic populations. (a) Illustrates the TSNE plot of real and fake data generated at the first epoch. (b) Illustrates the TSNE plot of real and fake data generated at last epoch

Chapter 5

Conclusion and Future Work

5.1 Summary of Contribution

Our research has been focused on designing generalizable deep learning (DL) models capable of learning continuously and adapting to data from diverse healthcare institutions. A key objective of our work is to ensure these models uphold the strict privacy standards required for protecting patient data. We designed new algorithms, and have conducted comparative analyses of various approaches, we focus our evaluation on specific biomedical researchers problem for predicting the onset of sepsis.

In our pursuit of developing generalizable deep learning predictive models, we have developed a new continual learning method that enable models to learn from ongoing data streams while retaining proficiency in previous tasks. In particular, We introduced an innovative continual learning algorithm, WUPERR, which utilizes weight consolidation and representation replay to help the model preserve acquired knowledge while learn new data. WUPERR specifically penalizes significant modifications to weights that are crucial for performing prior tasks, thereby mitigating the issue of catastrophic forgetting. Moreover, it ensures privacy by transferring only data representations between institutions. Comparative studies have demonstrated that our WUPERR model outperformed traditional transfer learning methods.

We then leveraged Federated Learning (FL) techniques, which facilitate collaborative learning while maintaining data privacy. In FL, data remains on-site at each healthcare institution where models are trained locally. These local models' weights are sent to a central server, ensuring patient privacy. The server then updates a global model weights with defined aggregation function and redistributes global models weights to the individual sites. We developed an innovative approach for updating the global model, which we have benchmarked against traditional federated learning aggregation functions. Our method has demonstrated superior performance compared to conventional FL, particularly in scenarios where the data distribution varies greatly across institutions.

Additionally, we utilized Generative Adversarial Networks (GANs) to create synthetic Electronic Health Records (EHR) data. Previous research in this field faced numerous challenges, which we addressed with a cutting-edge framework adopted from EHR-SAFE. Our framework boasts several benefits: it adeptly manages varying lengths of EHR sequences, employs stochastic normalization to mitigate skewed distributions, and utilizes Wasserstein loss with a gradient penalty to prevent model collapse. Our model has demonstrated promising results, particularly when using synthetic EHR data for continual learning. By replaying synthetic data, our model can retain previous knowledge efficiently, thus eliminating the need to transfer and replay real data, which addresses potential privacy concerns.

5.2 Future Direction

In our feature research work, we will focus on improving the generalizability and robustness of DL predictive models across varied clinical landscapes. Our objective includes the rapid adaptation of these models to diverse clinical settings and patient populations. A recognized limitation within our current framework is its susceptibility to outlier data influences in data representation and in turn model performance. Addressing this challenge,

we plan to develop and integrate more effective techniques for the detection and exclusion of outliers. This will involve creating sophisticated algorithms that can accurately identify and manage anomalies within each patient cohort at every participating healthcare institute. Our goal is to ensure that our model is trained on the most representative and accurate data, thus enhancing its performance and reliability in real-world clinical applications.

In our subsequent research initiatives, we will direct our focus towards the pivotal role of high-quality data and accurate labeling in achieving optimal model performance. We have observed that noise in the training data can significantly degrade the effectiveness of our models. Specifically, within our Federated Learning framework, it was noted that the model parameters fine-tuned using Hospital-C's data exhibited sub-optimal performance when tested on data from other hospitals. This challenge could be attributed to either a shift in data distribution domains or the presence of noisy labels. To tackle this issue, our future efforts will be towards leveraging advanced label enhancement techniques, with a particular emphasis on Graph Neural Networks (GNNs). GNNs operate under the premise that neighboring nodes tend to share similar labels. Utilizing this principle, we aim to construct GNNs within each institute by focusing on data points with high-quality labels, identifiable through lower cross-entropy loss. This strategy is expected to significantly improve the quality of labels, thereby enhancing the overall performance and reliability of our deep learning models across different healthcare settings. By refining the accuracy of our labels, we anticipate a substantial improvement in model robustness and its applicability to diverse clinical data.

Scaling the generation and application of synthetic EHR data is another critical area of focus. Our GAN framework needs to be optimized. we also need to adopt our model for longer sequence length and more complex data such as clinical notes, ensuring the synthetic data's utility is validated across a multitude of clinical prediction tasks. Rigorous testing for model resilience against data perturbations and ethical considerations will be

paramount, ensuring that the use of synthetic data and AI respects patient privacy and aligns with healthcare ethics.

Finally, integrating AI models into clinical practice necessitates close collaboration with healthcare professionals and ongoing real-world evaluation. Longitudinal studies will assess the impact on patient outcomes, guiding iterative model refinement. We will engage with regulatory bodies to ensure compliance and actively participate in shaping policies for AI in healthcare, fostering interdisciplinary collaborations that address the broad implications of our work. Through these efforts, we aim to create a future where AI-driven tools are seamlessly integrated into healthcare systems, enhancing the quality and accessibility of patient care.

5.3 Conclusion

In conclusion, our research has made significant contributions in the development of deep learning models that are both generalizable across different healthcare institutions and respectful of patient data privacy. We have evaluated different approaches for predicting sepsis, highlighting the strengths of our methodologies. Using continual learning algorithm, we have shown that models can effectively retain knowledge from sequential data streams without the common problem of catastrophic forgetting.

Moreover, our innovative use of Federated Learning has allowed for collaborative model training while keeping the data localized, ensuring patient privacy. We introduced a unique approach to updating the global model that surpasses traditional federated learning techniques, particularly in settings with disparate data distributions.

Lastly, the application of GANs for generating synthetic EHR data has addressed limitations found in prior studies. Our framework is specially designed to handle the challenges of dynamic EHR sequences and skewed data distributions. The synthetic data generated

has proven to be a valuable asset for continual learning processes, facilitating knowledge preserving without compromising data privacy.

These contributions mark a considerable advancement in the field, opening pathways to improved predictive models in healthcare while strictly adhering to privacy standards.

Bibliography

[1] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C. Folk, and Meindert Niemeijer. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, October 2016.

[2] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, September 2018. Number: 9 Publisher: Nature Publishing Group.

[3] Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, June 2020.

[4] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020.

[5] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731, October 2018.

[6] Yadi Zhou, Fei Wang, Jian Tang, Ruth Nussinov, and Feixiong Cheng. Artificial intelligence in COVID-19 drug repurposing. *The Lancet Digital Health*, 2(12):e667–e676, December 2020. Publisher: Elsevier.

[7] Jenny Yang, Andrew A. S. Soltan, and David A. Clifton. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *NPJ digital medicine*, 5(1):69, June 2022.

[8] The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information | paper by Miller | Britannica.

[9] The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) | Critical Care Medicine | JAMA | JAMA Network.

[10] Derek C. Angus and Tom van der Poll. Severe sepsis and septic shock. *The New England Journal of Medicine*, 369(9):840–851, August 2013.

[11] Celeste M. Torio and Roxanne M. Andrews. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011. In *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*. Agency for Healthcare Research and Quality (US), Rockville (MD), 2006.

[12] Carly J. Paoli, Mark A. Reynolds, Meenal Sinha, Matthew Gitlin, and Elliott Crouser. Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Critical Care Medicine*, 46(12):1889–1897, December 2018.

[13] Chanu Rhee, Raymund Dantes, Lauren Epstein, David J. Murphy, Christopher W. Seymour, Theodore J. Iwashyna, Sameer S. Kadri, Derek C. Angus, Robert L. Danner, Anthony E. Fiore, John A. Jernigan, Greg S. Martin, Edward Septimus, David K. Warren, Anita Karcz, Christina Chan, John T. Menchaca, Rui Wang, Susan Gruber, Michael Klompas, and CDC Prevention Epicenter Program. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA*, 318(13):1241–1249, October 2017.

[14] The Surviving Sepsis Campaign Guidelines Committee including The Pediatric Subgroup*, R. P. Dellinger, Mitchell M. Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M. Opal, Jonathan E. Sevransky, Charles L. Sprung, Ivor S. Douglas, Roman Jaeschke, Tiffany M. Osborn, Mark E. Nunnally, Sean R. Townsend, Konrad Reinhart, Ruth M. Kleinpell, Derek C. Angus, Clifford S. Deutschman, Flavia R. Machado, Gordon D. Rubenfeld, Steven Webb, Richard J. Beale, Jean-Louis Vincent, and Rui Moreno. Surviving Sepsis Campaign: International Guidelines for Management of Severe Sepsis and Septic Shock, 2012. *Intensive Care Medicine*, 39(2):165–228, February 2013.

[15] Arjun K. Venkatesh, Umakanth Avula, Holly Bartimus, Justin Reif, Michael J. Schmidt, and Emilie S. Powell. Time to antibiotics for septic shock: evaluating a proposed performance measure. *The American Journal of Emergency Medicine*, 31(4):680–683, April 2013.

[16] Sarah A. Sterling, W. Ryan Miller, Jason Pryor, Michael A. Puskarich, and Alan E. Jones. The Impact of Timing of Antibiotics on Outcomes in Severe Sepsis and Septic Shock: A Systematic Review and Meta-Analysis. *Critical Care Medicine*, 43(9):1907–1915, September 2015.

[17] Andrew Rhodes, Gary Phillips, Richard Beale, Maurizio Cecconi, Jean Daniel Chiche, Daniel De Backer, Jigeeshu Divatia, Bin Du, Laura Evans, Ricard Ferrer, Massimo Girardis, Despoina Koulenti, Flavia Machado, Steven Q. Simpson, Cheng Cheng Tan, Xavier Wittebole, and Mitchell Levy. The Surviving Sepsis Campaign bundles and outcome: results from the International Multicentre Prevalence Study on Sepsis (the IMPreSS study). *Intensive Care Medicine*, 41(9):1620–1628, September 2015.

[18] Ricard Ferrer, Ignacio Martin-Loeches, Gary Phillips, Tiffany M. Osborn, Sean Townsend, R. Phillip Dellinger, Antonio Artigas, Christa Schorr, and Mitchell M. Levy. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Critical Care Medicine*, 42(8):1749–1755, August 2014.

[19] Mitchell M. Levy, Laura E. Evans, and Andrew Rhodes. The Surviving Sepsis Campaign Bundle: 2018 update. *Intensive Care Medicine*, 44(6):925–928, June 2018.

[20] Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *The New England Journal of Medicine*, 376(23):2235–2244, June 2017.

[21] Andrew F. Shorr, Scott T. Micek, William L. Jackson, and Marin H. Kollef. Economic implications of an evidence-based sepsis protocol: can we improve outcomes and lower costs? *Critical Care Medicine*, 35(5):1257–1262, May 2007.

[22] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest*, 101(6):1644–1655, June 1992.

[23] Bryan Williams. The National Early Warning Score: from concept to NHS implementation. *Clinical Medicine*, 22(6):499–505, November 2022. Publisher: Royal College of Physicians Section: 10 Years of news.

[24] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical care medicine*, 46(4):547–553, April 2018.

[25] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier, June 2017. arXiv:1706.04152 [stat].

[26] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D. Feldman, Chris Barton, David J. Wales, and Ritankar Das. Prediction of Sepsis in the Intensive Care Unit

With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR medical informatics*, 4(3):e28, September 2016.

[27] Supreeth P. Shashikumar, Christopher Josef, Ashish Sharma, and Shamim Nemati. DeepAISE – An Interpretable and Recurrent Neural Survival Model for Early Prediction of Sepsis. *Artificial intelligence in medicine*, 113:102036, March 2021.

[28] Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122, August 2015.

[29] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA internal medicine*, 181(8):1065–1070, August 2021.

[30] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

[31] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience Replay for Continual Learning, November 2019. arXiv:1811.11682 [cs, stat].

[32] Supreeth P. Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. Artificial Intelligence Sepsis Prediction Algorithm Learns to Say "I don't know", May 2021. ISSN: 2125-6764 Pages: 2021.05.06.21256764.

[33] Bogdan Draganski, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Bogdahn, and Arne May. Changes in grey matter induced by training. *Nature*, 427(6972):311–312, January 2004. Number: 6972 Publisher: Nature Publishing Group.

[34] Claudia Clopath. Synaptic consolidation: an approach to long-term learning. *Cognitive Neurodynamics*, 6(3):251–257, June 2012.

[35] David Meunier, Renaud Lambiotte, and Edward T. Bullmore. Modular and Hierarchically Modular Organization of Brain Networks. *Frontiers in Neuroscience*, 4:200, December 2010.

[36] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning, December 2013. arXiv:1312.5602 [cs].

[37] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-End Incremental Learning. pages 233–248, 2018.

[38] ANTHONY ROBINS. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connection Science*, 7(2):123–146, June 1995. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09540099550039318.

[39] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3987–3995. PMLR, July 2017. ISSN: 2640-3498.

[40] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational Continual Learning, May 2018. arXiv:1710.10628 [cs, stat].

[41] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks, October 2022. arXiv:1606.04671 [cs].

[42] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong Learning with Dynamically Expandable Networks, June 2018. arXiv:1708.01547 [cs].

[43] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural Architecture Search: A Survey, April 2019. arXiv:1808.05377 [cs, stat].

[44] Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & Compress: A scalable framework for continual learning, July 2018. arXiv:1805.06370 [cs, stat].

[45] Stephen James, Michael Bloesch, and Andrew J. Davison. Task-Embedded Control Networks for Few-Shot Imitation Learning, October 2018. arXiv:1810.03237 [cs].

[46] Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional Neural Processes, July 2018. arXiv:1807.01613 [cs, stat].

[47] Dani Kiyasseh, Tingting Zhu, and David Clifton. A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions. *Nature Communications*, 12:4221, July 2021.

[48] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[49] Supreeth Prajwal Shashikumar. *Generalizable Models for Prediction of Physiological Decompensation from multivariate and Multiscale Physiological Time Series using Deep Learning and Transfer Learning Techniques.* PhD Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2021.

[50] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, February 2016.

[51] Gabriel Wardi, Morgan Carlile, Andre Holder, Supreeth Shashikumar, Stephen R. Hayden, and Shamim Nemati. Predicting Progression to Septic Shock in the Emergency Department Using an Externally Generalizable Machine-Learning Algorithm. *Annals of Emergency Medicine*, 77(4):395–406, April 2021.

[52] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1814259.

[53] Beau Norgeot, Benjamin S. Glicksberg, and Atul J. Butte. A call for deep-learning healthcare. *Nature Medicine*, 25(1):14–15, January 2019. Number: 1 Publisher: Nature Publishing Group.

[54] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, November 2018.

[55] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):1–7, September 2020. Number: 1 Publisher: Nature Publishing Group.

[56] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated Learning for Healthcare Informatics, August 2020. arXiv:1911.06270 [cs].

[57] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):12598, July 2020. Number: 1 Publisher: Nature Publishing Group.

[58] Li Huang, Andrew L. Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of Biomedical Informatics*, 99:103291, November 2019.

[59] Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir. Collaborative Federated Learning For Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge, January 2021. arXiv:2101.07511 [cs].

[60] Andrew Wong, Jie Cao, Patrick G. Lyons, Sayon Dutta, Vincent J. Major, Erkin Ötleş, and Karandeep Singh. Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic. *JAMA Network Open*, 4(11):e2135286, November 2021.

[61] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, Picking and Growing for Unforgetting Continual Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[62] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic For-

getting. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3925–3934. PMLR, May 2019. ISSN: 2640-3498.

[63] Andre L Holder, Supreeth P Shashikumar, Gabriel Wardi, Timothy G Buchman, and Shamim Nemati. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the icu. *Critical care medicine*, 49(12):e1196–e1205, 2021.

[64] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):1–7, September 2020. Number: 1 Publisher: Nature Publishing Group.

[65] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, January 2023. arXiv:1602.05629 [cs].

[66] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency, October 2017. arXiv:1610.05492 [cs].

[67] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, March 2021.

[68] Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. Federated Learning for Breast Density Classification: A Real-World Implementation. volume 12444, pages 181–191. 2020. arXiv:2009.01871 [cs, eess].

[69] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, and Andrew Feng. Privacy-preserving Federated Brain Tumour Segmentation, October 2019. arXiv:1910.00962 [cs].

[70] Hyunghoon Cho, David Froelicher, Jeffrey Chen, Manaswitha Edupalli, Apostolos Pyrgelis, Juan R. Troncoso-Pastoriza, Jean-Pierre Hubaux, and Bonnie Berger. Secure and Federated Genome-Wide Association Studies for Biobank-Scale Datasets, December 2022. Pages: 2022.11.30.518537 Section: New Results.

[71] Ittai Dayan, Holger R. Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z. Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J. Wood, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, C. K. Lee, Peiying Ruan, Daguang Xu,

Dufan Wu, Eddie Huang, Felipe Campos Kitamura, Griffin Lacey, Gustavo César de Antônio Corradi, Gustavo Nino, Hao-Hsin Shin, Hirofumi Obinata, Hui Ren, Jason C. Crane, Jesse Tetreault, Jiahui Guan, John W. Garrett, Joshua D. Kaggie, Jung Gil Park, Keith Dreyer, Krishna Juluru, Kristopher Kersten, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena AbdelMaseeh, Nicola Rieke, Pablo F. Damasceno, Pedro Mario Cruz e Silva, Pochuan Wang, Sheng Xu, Shuichi Kawano, Sira Sriswasdi, Soo Young Park, Thomas M. Grist, Varun Buch, Watsamon Jantarabenjakul, Weichung Wang, Won Young Tak, Xiang Li, Xihong Lin, Young Joon Kwon, Abood Quraini, Andrew Feng, Andrew N. Priest, Baris Turkbey, Benjamin Glicksberg, Bernardo Bizzo, Byung Seok Kim, Carlos Tor-Díez, Chia-Cheng Lee, Chia-Jung Hsu, Chin Lin, Chiu-Ling Lai, Christopher P. Hess, Colin Compas, Deepeksha Bhatia, Eric K. Oermann, Evan Leibovitz, Hisashi Sasaki, Hitoshi Mori, Isaac Yang, Jae Ho Sohn, Krishna Nand Keshava Murthy, Li-Chen Fu, Matheus Ribeiro Furtado de Mendonça, Mike Fralick, Min Kyu Kang, Mohammad Adil, Natalie Gangai, Peerapon Vateekul, Pierre Elnajjar, Sarah Hickman, Sharmila Majumdar, Shelley L. McLeod, Sheridan Reed, Stefan Gräf, Stephanie Harmon, Tatsuya Kodama, Thanyawee Puthanakit, Tony Mazzulli, Vitor Lima de Lavor, Yothin Rakvongthai, Yu Rim Lee, Yuhong Wen, Fiona J. Gilbert, Mona G. Flores, and Quanzheng Li. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10):1735–1743, October 2021. Number: 10 Publisher: Nature Publishing Group.

[72] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated Multi-Task Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[73] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated Learning: Challenges, Methods, and Future Directions, August 2019. arXiv:1908.07873 [cs, stat].

[74] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, January 2023. arXiv:1602.05629 [cs].

[75] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[76] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. Secure Federated Transfer Learning. *IEEE Intelligent Systems*, 35(4):70–82, July 2020. arXiv:1812.03337 [cs, stat].

[77] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes (Workshop)*, 11383:92–104, 2019.

[78] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics*, 112:59–67, April 2018.

[79] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[80] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, January 2016. arXiv:1511.06434 [cs].

[81] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, December 2017. arXiv:1701.07875 [cs, stat].

[82] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs, June 2016. arXiv:1606.03498 [cs].

[83] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs, July 2017. arXiv:1610.09585 [cs, stat].

[84] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation, February 2018. arXiv:1710.10196 [cs, stat].

[85] Kong, J., Kim, J. & Bae, J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural Inf. Process. Syst. 33, 17022–17033 (2020). - Google Search.

[86] Cyprien de Masson d' Autume, Shakir Mohamed, Mihaela Rosca, and Jack Rae. Training Language GANs from Scratch. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[87] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[88] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks, January 2018. arXiv:1703.06490 [cs].

[89] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs, December 2017. arXiv:1706.02633 [cs, stat].

[90] Brett K. Beaulieu-Jones and Jason H. Moore. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:207–218, 2017.

[91] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S. Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, Farhana Bandukwala, Elli Kanal, Sercan Ö Arık, and Tomas Pfister. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine*, 6(1):1–11, August 2023. Number: 1 Publisher: Nature Publishing Group.

[92] All of Us Research Program Investigators, Joshua C. Denny, Joni L. Rutter, David B. Goldstein, Anthony Philippakis, Jordan W. Smoller, Gwynne Jenkins, and Eric Dishman. The "All of Us" Research Program. *The New England Journal of Medicine*, 381(7):668–676, August 2019.

[93] Fatemeh Amrollahi, Supreeth P Shashikumar, Angela Meier, Lucila Ohno-Machado, Shamim Nemati, and Gabriel Wardi. Inclusion of social determinants of health improves sepsis readmission prediction models. *Journal of the American Medical Informatics Association : JAMIA*, 29(7):1263–1270, May 2022.

[94] Divya Saxena and Jiannong Cao. Generative Adversarial Networks (GANs Survey): Challenges, Solutions, and Future Directions, April 2023. arXiv:2005.00065 [cs, eess, stat].

[95] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[96] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. arXiv:1706.08500 [cs, stat].

[97] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks, March 2019. arXiv:1812.04948 [cs, stat].

[98] Olof Mogren. C-RNN-GAN: Continuous recurrent neural networks with adversarial training, November 2016. arXiv:1611.09904 [cs].