TEMPORALLY-EMBEDDED DEEP LEARNING FOR HEALTH PREDICTION

TEMPORALLY-EMBEDDED DEEP LEARNING MODEL FOR HEALTH OUTCOME PREDICTION

BY

OMAR BOURSALIE, B.Eng., M.A.Sc.

A THESIS

SUBMITTED TO THE SCHOOL OF BIOMEDICAL ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2021) McMaster University

(School of Biomedical Engineering) Hamilton, Ontario, Canada

TITLE:	Temporally-Embedded Deep Learning Model for Health Outcome Prediction
AUTHOR:	Omar Boursalie B.Eng. M.A.Sc., (Biomedical Engineering) McMaster University, Hamilton, Canada
SUPERVISORS:	Dr. Thomas Doyle and Dr. Reza Samavi
NUMBER OF PAGES:	xvii, 116

Lay Abstract

In this thesis, two challenges using deep learning models to analyze health records are investigated using a real-world medical dataset. First, an important step in analyzing health records is to estimate missing data. We investigated how imputation can have a cascading negative impact on a deep learning model's performance. A comparative analysis was then conducted to investigate the strengths and limitations of evaluation metrics from the statistical literature to assess deep learning-based imputation models. Second, the most successful deep learning diagnostic models to date, called transformers, lack a mechanism to analyze the temporal characteristics of health records. To address this gap, we developed a new temporally-embedded transformer to analyze patients' medical histories, including the elapsed time between visits, to predict their primary diagnoses. The proposed model successfully predicted patients' primary diagnosis in their final visit with improved predictive performance ($78.54 \pm 0.22\%$) compared to existing models in the literature.

Abstract

Deep learning models are increasingly used to analyze health records to model disease progression. Two characteristics of health records present challenges to developers of deep learning-based medical systems. First, the veracity of the estimation of missing health data must be evaluated to optimize the performance of deep learning models. Second, the currently most successful deep learning diagnostic models, called transformers, lack a mechanism to analyze the temporal characteristics of health records.

In this thesis, these two challenges are investigated using a real-world medical dataset of longitudinal health records from 340,143 patients over ten years called MI-IDD: McMaster Imaging Information and Diagnostic Dataset. To address missing data, the performance of imputation models (mean, regression, and deep learning) were evaluated on a real-world medical dataset. Next, techniques from adversarial machine learning were used to demonstrate how imputation can have a cascading negative impact on a deep learning model. Then, the strengths and limitations of evaluation metrics from the statistical literature (qualitative, predictive accuracy, and statistical distance) to evaluate deep learning-based imputation models were investigated. This research can serve as a reference to researchers evaluating the impact of imputation on their deep learning models.

To analyze the temporal characteristics of health records, a new model was developed and evaluated called DTTHRE: Decoder Transformer for Temporally-Embedded Health Records Encoding. DTTHRE predicts patients' primary diagnoses by analyzing their medical histories, including the elapsed time between visits. The proposed model successfully predicted patients' primary diagnosis in their final visit with improved predictive performance ($78.54 \pm 0.22\%$) compared to existing models in the literature. DTTHRE also increased the training examples available from limited medical datasets by predicting the primary diagnosis for each visit ($79.53 \pm 0.25\%$) with no additional training time. This research contributes towards the goal of disease predictive modeling for clinical decision support.

Acknowledgements

It is my great pleasure and honour to thank everyone who supported me to develop this thesis, although it is possible to give particular mention only to a few here.

I sincerely thank my supervisors, Dr. Thomas E. Doyle and Dr. Reza Samavi for their confidence, trust, patience, guidance, support, passion, critical suggestions, and discussions throughout my thesis. It has been an honour and privilege to have the opportunity to learn so much from both of you.

I also warmly thank the members of my thesis committee. Dr. Hubert de Bruin and Dr. Michael Noseworthy who provided detailed guidance that greatly influenced my research. Dr. David A. Koff who generously offered his time to help me understand the medical aspects of my research. Dr. Samantha Kleinberg, the external examiner, who appraised my work and provided valuable comments that improved my thesis. I also thank the anonymous reviewers whose suggestions improved and clarified my papers presented in Chapters 3 - 5.

I also acknowledge Hamilton Health Sciences (HHS) for contributing the medical dataset used in this thesis. A special thanks to Jane A. Castelli, our project manager, who worked tirelessly with Stephen Melnyk and Vasilic Lillian (HHS) to construct and access the medical dataset. I also thank Ken Mascola (HHS) who helped me understand the technical details of the medical imaging scanners studied in this thesis.

I also thank all my labmates in the Biomedic.AI Lab and The Trustworthy Artificial Intelligence Research Lab (TAILab) for creating a warm and supportive environment to grow and learn together. A special thanks to Andrew Sutton for his support with the medical dataset used in this thesis.

I also thank my colleagues and friends Devin Packer, George Su, Krystien and Nicky Wolf, Avery Chakravarty, and Jessica Anderson who provided constant support throughout my studies. A special thanks to Micheal Wirtzfeld and Geneva Smith for their support and writing advice. You have all enriched my graduate experience.

Outside of McMaster, I could always count on the support of my family and friends: Mom, Dad, Suz, Rob, Jess, Calvin and Renée Bouwman, Jackson and Libby Ploeg, and Joshua and Angela Tim. Without your never-ending support, I could not have finished this thesis.

This thesis was supported by McMaster University, Natural Sciences and Engineering Research Council of Canada (NSERC), Vector Institute for Artificial Intelligence, Southern Ontario Smart Computing Innovation Platform (SOSCIP), and Canadian Department of National Defence: Innovation for Defence Excellence and Security (IDEaS) Program.

Lastly, I thank all of my students for starting me on this journey.

To my family

Contents

List of Abbreviations and Symbols

Abbreviations

BERT Bidirectional Encoder Representations from Transformers
CDT Cohen's Distance Test
CIHI Canadian Institute for Health Information
CNN Convolutional Neural Network
CT Computed Tomography
$CTDI_{vol}$ Computed Tomography Dose Index Volume
CV Cross-validation
DAD Discharge Abstract Database
DAE Denoising Autoencoder
DICOM Digital Imaging and Communications in Medicine
DSD Distance Source to Detector
DSS Decision Support System
DTTHRE Decoder Transformer for Temporally-Embedded Health Records Encoding
Dx Diagnosis
ED Effective Dose
EU European Union
GAIN Generative Adversarial Imputation Nets
GAN Generative Adversarial Nets
GE General Electric
GPT Generative Pre-trained Transformer
Gy Gray
HAM Hamilton
HHS Hamilton Health Sciences
ICD International Classification of Diseases
ICRP International Commission on Radiological Protection
JSDist Jensen Shannon Distance
KL Kullback-Leibler
kVp X-ray Tube Voltage
LNT Linear-No-Threshold
LSTM Long Short-Term Memory
MAR Missing at Random
MCAR Missing Completely at Random

MIDAS Multiple Imputation with Denoising Autoencoders
MIDI Musical Instrument Digital Interface
MIIDD McMaster Imaging Information and Diagnostic Dataset
MIMIC Medical Information Mart for Intensive Care
MLM Masked Language Modeling
MNAR Missing Not at Random
NACRS National Ambulatory Care Reporting System
NCICT National Cancer Institute Dosimetry System for Computed Tomography
NLP Natural Language Processing
PACS Picture Archiving and Communications Systems
PMM Predictive Mean Matching
QT Quantile Transform
RMSE Root Mean Square Error
RNN Recurrent Neural Network
SE Standard Error
SEL Scan End Location
SPF Spiral Pitch Factor
SSL Scan Start Location
Sv Sievert
SVM Support Vector Machine
TCW Total Collimation Width
THRE Temporally-Embedded Health Records Encoding
VAE Variational Autoencoders
XR X-ray

Symbols

Δt Elapsed time between subsequent visits
δ Perturbation
$\hat{b}$ Imputed values
$\hat{Q}$ Population statistics
λ Distance between subsequent sequence elements
ϕ Divergence metrics
ψ Probability of data being missing
τ Exposure time
$\mathbf{X*}$ Adversarial sample
$\mathbf{X}$ Input feature vector
a Age
AE Autoencoder
b Actual values
C Total effective dose
c Draws from distribution
D Matrix of training data
d Diagnostic sequence
$D_{(0)}$ Missing data

$D_{(1)}$	Observed data
DC	Matrix of Monte Carlo calculations
e	Imaging exam type
f	Feature
f_c	Feed-forward classification layer
f_d	Decoder layer
f_e	Encoder layer
G	Gender-specific equivalent dose
H	Patient medical history (irregular time series)
h	Patient medical history (sequence)
I	Tube current
i	Sequence element
j	Final element in sequence
k	Fold
k_{OB}	Overbeaming correction factor
l	Label
m	Medicine or treatment sequence
N	Total number of imaging scans
n_p	Number of visits per patient p
$nCTDI_w$	Volume computed tomography dose index
P	Patient
PC	Percentage change
Q	Population parameters
R	Response matrix
r	Instance
s	Sex
T	Organ or tissue
t	Observation time
u	Scan slice number
v	Patient visit
w	Weighting factors
w_R	International Commission on Radiological Protection radidation weighting factor
x	Scan
y	Prediction
z	Continuous representation

List of Figures

List of Tables

Declaration of Academic Achievement

The following is a declaration that the research described in this thesis was completed by Mr. Omar Boursalie and recognizes the contributions of Dr. Thomas Doyle and Dr. Reza Samavi. Omar Boursalie contributed to the inception and design of the study. Omar Boursalie was also responsible for the experimental testing protocols, design and development of the experiments and models, data collection, data analysis, and the writing of the manuscript. Dr. Thomas Doyle and Dr. Reza Samavi contributed to the inception and design of the study and the review of the manuscript.

Chapter 1

Introduction

Deep learning models are increasingly being used to predict patients' diagnoses by analyzing the trends and relationships in their medical histories. The deep learning models are generally designed to analyze datasets such as text and images that are from a single source (homogeneous) with no missing data and no time component (non-temporal). In contrast, medical histories recorded in electronic health records (EHR) have the following characteristics. First, modeling medical histories involve analyzing data from multiple sources (heterogeneous) such as diagnostic, physiological, and imaging. As a result, missing data is common due to the technical and privacy challenges of collecting health data from multiple sources. Second, medical records represent multiple observations of a patient's health over time (temporal). The following real-world health scenario demonstrates the challenges of analyzing heterogeneous and temporal medical datasets with missing data using deep learning.

As part of an interdisciplinary team at McMaster, Ryerson University, and Hamilton Health Sciences (HHS), I am currently developing a decision support system

(DSS) using deep learning that provides real-time risk assessment of radiation exposure from medical imaging relative to a patient's medical history. Two challenges developing the risk model were encountered: 1) missing data in health records and 2) analyzing the temporal characteristics of medical histories. First, an important step towards developing the DSS was collecting a dataset of real medical histories for training and evaluating the models called MIIDD: McMaster Imaging Information and Diagnostic Dataset. Due to technical and privacy challenges, I accessed a subset of patients' heterogeneous records to estimate exposure from medical imaging. As a result, the low-dose radiation exposure for the remaining images in MIIDD must be imputed. Imputation is the process of replacing missing data with estimated values. An important task in imputing data is evaluating the performance of the imputation models (Nguyen, Carlin, & Lee, 2017). Second, the risk model must analyze patients' exposure patterns over time within the context of their medical histories. Deep learning models generally lack a mechanism to analyze temporal patterns in EHR. The following sections introduce the background of the research, outline the thesis, and summarize the main contributions.

1.1 Imputation of Missing Data in Health Records

Figure 1.1 shows an overview of the research presented in this thesis based on the methodology used to develop deep learning models. The first step towards developing the deep learning model for health outcome prediction was curating MIIDD (Fig. 1.1a). MIIDD contains approximately 2.1 million imaging records from 340,143 patients over ten years in four hospitals in Hamilton, Ontario, Canada. The medical records are stored in three repositories: 1) health records of diagnostic codes, 2)

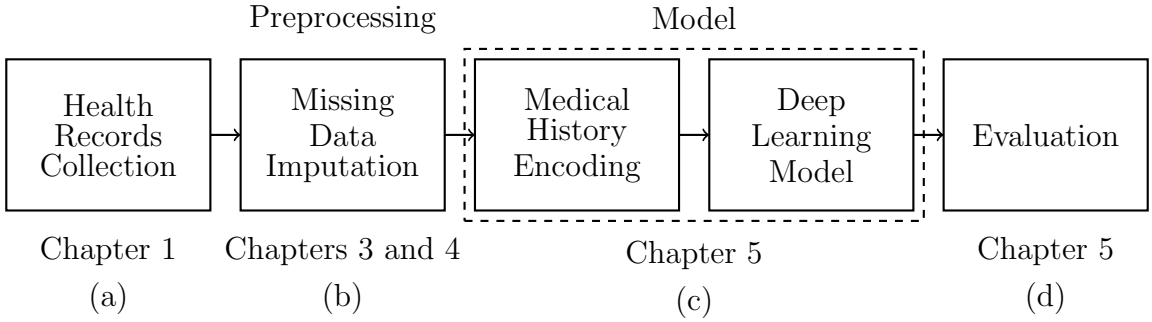


Figure 1.1: Thesis overview based on the development methodology for deep learning models. Literature Review and Conclusions are in Chapters 2 and 6, respectively

imaging records that contain summary data on scans (e.g., modality, body part, and date of scan), and 3) diagnostic images that can provide a more accurate estimate of exposure. Due to technical and privacy challenges, a representative subset of diagnostic images was accessed to estimate exposure. As a result, the exposure for the remaining imaging records must be imputed. The imputation model's performance was evaluated as it can have a cascading impact on the deep learning risk model.

The investigation evaluating imputation models for health records led to three studies (Fig. 1.1b). First, the performance of mean, regression, and deep learning imputation models were evaluated to estimate missing data in heterogeneous health records for a target feature (imaging exposure) over a patient's medical history. Second, there is a need to better understand the impact of imprecise imputation methods on deep learning performance. To that end, I used techniques from adversarial machine learning (Papernot et al., 2016) to investigate how mean imputation impacts the performance of a proof-of-concept deep learning model. Third, discrepancies were discovered evaluating deep learning-based imputation models using the common evaluation metric used in the literature, root mean square error (RMSE), compared to qualitative evaluation. Investigating this discrepancy led to a comparative analysis

of evaluation metrics from the statistical literature (qualitative, predictive accuracy, and statistical distance) to assess the performance of deep learning-based imputation models. Based on the findings, the missing exposure values in MIIDD were estimated using the imputation model that performed best on the representative sample. The next step was to model patients' medical histories represented by the imputed dataset using deep learning.

1.2 Temporally-Embedded Deep Learning

The second research objective in this thesis was to model the temporal characteristics of health records using deep learning (Fig. 1.1c). A common approach (Y. Li et al., 2020; Rasmy, Xiang, Xie, Tao, & Zhi, 2021) for analyzing the temporal patterns in medical histories is to encode them as diagnostic sequences (diagnostic-level encoding). A sequence is a set of elements (e.g., diagnoses) that are listed in order. Deep learning models such as recurrent neural networks (RNN; Rumelhart, Hinton, and Williams, 1985) and long-short term memory models (LSTM; Hochreiter and Schmidhuber, 1997) analyze the sequential order of medical diagnoses to learn disease patterns (Shickel, Tighe, Bihorac, & Rashidi, 2018). RNN and LSTM require sequential processing, and the information from previous visits is stored recursively in the model's memory.

Transformers (Vaswani et al., 2017) are a class of deep learning models that learn to relate elements in a sequence to each other by generating weight-adjusted representations for each sequence. The amount of weight (attention) applied to each sequence element is learned by the model during training. The encoded representation of the sequence is then used by the transformer to predict elements in the

sequence. Transformers have three advantages over the RNN and LSTM (Vaswani et al., 2017). First, transformers access sequence elements directly rather than through the recursive memory used in RNN and LSTM. Second, transformers analyze the elements in each sequence at once, which enables parallelization to reduce training time. Third, transformer models can be trained on larger datasets compared to RNN and LSTM. The transformer has demonstrated improved performance for natural language processing (NLP; Vaswani et al., 2017) and diagnostic prediction (Y. Li et al., 2020) tasks compared to RNN and LSTM.

An important requirement for assessing patients' risk from medical imaging is analyzing the elapsed time between exposure events. Transformers are designed for NLP where the distance between subsequent elements is constant. In contrast, patients' visits to their health professionals are episodic and the time between subsequent visits varies (irregular time series). For example, consider two patients who have the same low-dose radiation exposure from three medical scans. The first patient has three scans over two months while the second patient has the scans over two years. Transformer models lack a mechanism for analyzing the elapsed time between elements in a sequence. As a result, a transformer analyzing exposure histories would represent the patients' exposure patterns with the same sequence. There is a need to develop a transformer model that analyzes the elapsed times between visits when predicting disease trajectories.

In this thesis, a transformer model called Decoder Transformer for Temporally-Embedded Health Records Encoding (DTTHRE) was developed that predicts the primary diagnosis (Dx) for each visit using the patient's medical history, including the elapsed times between visits (Fig. 1.1c). The proposed model requires an encoding

mechanism that embeds the irregular time series data in medical histories. Instead of diagnostic-level encoding, an encoding representation for EHR was proposed called Temporally-Embedded Health Records Encoding (THRE). THRE encodes EHR as sequences of medical events such as age, sex, and visit-level diagnostic embeddings while incorporating the elapsed time between visits. Instead of predicting each element in THRE, DTTHRE predicts the primary diagnosis for each visit by analyzing all events from previous medical visits.

A proof-of-concept DTTHRE model was developed to evaluate if embedding the time between visits impacts predictive performance (Fig. 1.1d). The DTTHRE's performance was then compared to an existing diagnostic model in the literature. DTTHRE successfully analyzed patients' medical histories, including the elapsed time between visits, to predict the primary diagnosis in their final visit with improved predictive performance ($78.54 \pm 0.22\%$) compared to the existing model.

1.3 Thesis Outline and Major Contributions

The thesis structure and contributions are as follows.

Chapter 2 presents the related work and gap analysis on the relevant research in three areas. The first area of the literature review investigates the requirements for exposure risk assessment models, motivating the investigation into missing data imputation and analyzing temporal data using deep learning. The second related research area reviews imputation models and their evaluation methodology. The third area of the literature review investigates transformer models for health record analysis.

Chapter 3 reports the investigation into the performance of mean imputation to

estimate missing data in health records and the impact on a deep learning model (Fig. 1.1b). I make the following contributions in this chapter:

1. A comparative analysis between two methods to estimate exposure from medical imaging: 1) dose means from the literature and 2) dose estimates calculated from real imaging scans.

2. Techniques from adversarial machine learning are used to investigate how mean imputation impacts the performance of a deep learning model.

Chapter 4 reports on a comparative study between evaluation metrics in the statistical literature to assess the performance of deep learning imputation models (Fig. 1.1b). I make the following contributions in this chapter:

1. A survey of the available evaluation metrics from the statistical literature (qualitative, predictive accuracy, and statistical distance) to evaluate deep learning imputation models.

2. A comparative analysis of the reviewed evaluation metrics to assess the performance of two deep learning-based imputations models and a regression imputation model using two heterogeneous datasets.

3. Investigate the strengths and limitations of using the evaluation metrics to assess the quality of deep learning imputation models.

Chapter 5 describes the proposed deep learning model architecture for analyzing patients' medical records, including the elapsed time between visits (Fig. 1.1c-d). I make the following contributions in this chapter:

1. DTTHRE, a transformer model that analyzes medical histories, including the elapsed time between visits, to predict diagnoses.

2. THRE, a temporally-embedded encoding representation for health records.

3. Develop a proof-of-concept DTTHRE and evaluate the model's performance compared to an existing transformer diagnostic model in the literature.

Chapter 6 presents conclusions and discusses directions for future work. First, the potential of defence mechanisms from the adversarial learning literature to improve the resilience of deep learning to imprecise imputation models are discussed. Second, the requirements for an evaluation methodology specifically for deep learning imputation models are described. Third, extensions to DTTHRE are proposed to investigate the relationships learned by the model and predict disease trajectories.

All evaluations in this thesis were done on a real medical records dataset I curated called the McMaster Imaging Information and Diagnostic Dataset (Fig. 1.1a). MIIDD contains longitudinal medical records covering ten years (May 2006 - 2016) from four hospitals in Hamilton, Ontario, Canada. The dataset characteristics and EHR samples are shown in Appendix A and B, respectively. MIIDD (Table B.1) contains deidentified, consolidated, cleaned, and linked health records from three medical repositories: 1) ambulatory data from the Discharge Abstract Database (DAD), 2) inpatient data from the National Ambulatory Care Reporting System (NACRS), and 3) imaging data from the HHS's Picture Archiving and Communications Systems (PACS). DAD (Table B.2) contains administrative, clinical, and demographic information on hospital discharges. NACRS (Table B.3) contains clinical information (e.g., diagnosis and treatments) from day surgery, outpatient, and community-based clinics.

Each hospital reports DAD and NACRS records to the Canadian Institute for Health Information (CIHI) each year to comply with Canadian law. The PACS (Table B.4) contains summary data on imaging scans (e.g., modality and body part scanned) and detailed imaging data stored in the DICOM (digital imaging and communications in medicine) format that contains the scanner data required to estimate patient exposure (Table B.5). The PACS system is used by all the hospitals to store their imaging data to comply with Canadian law. MIIDD has approximately 1.3 million diagnostic and 2.1 million imaging records from 340,143 patients. Due to technical and privacy challenges, I accessed a subset of 39,909 DICOM headers from a representative subset of 2,000 patients. MIIDD also includes the month and year of each medical visit.

The author brings to the reader's attention that the work in this thesis was published in Boursalie, Samavi, Doyle, and Koff, 2020b. In addition, work in this thesis has been accepted for publication in Boursalie, Samavi, and Doyle, 2021a and Boursalie, Samavi, and Doyle, 2021b. These publications were made by the author of this thesis, as the lead author, in collaboration with his supervisors at McMaster University. The review of the related literature and gap analysis in Chapter 2 are the contributions that have only been published in this thesis.

Chapter 2

Literature Review

In this chapter, a review of the related research is presented in three areas: 1) requirements for exposure risk assessment models, 2) imputation models and their evaluation methodology, and 3) transformers for health record analysis. These three areas of research are examined in Sections 2.1, 2.2, and 2.3, respectively.

2.1 Exposure Model Requirements

Multiple exposure models have been proposed to calculate a patient's risk of disease or mortality by analyzing their pattern of exposure (Fig. 2.1) to toxicants such as ionizing radiation (Lin, 2010), pollution (Vitolo, Scutari, Ghalaieny, Tucker, & Russell, 2018), or toxins (M. Liu et al., 2011). Current ionization radiation risk models on the effects of radiation exposure are based on statistics from the Japanese atomic bomb survivors (National Research Council, 2006). The Linear-No-Threshold (LNT) model linearly extrapolates the cancer risk assessment from high radiation doses to the low level of radiation emitted by imaging devices based on age and sex (Royal,

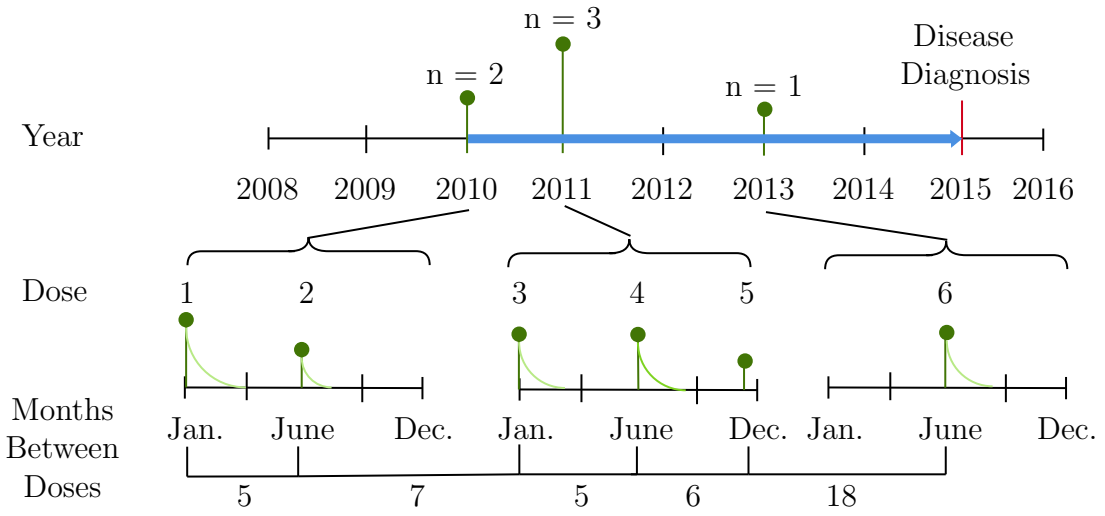


Figure 2.1: Timeline of a patient's exposure (green) and medical history (blue)

2008). Ionizing radiation risk assessment models such as RadRat (Gonzalez, Apostoaei, Veiga, & Land, 2013) then calculate a patients' lifetime and organ-specific cancer risk based on their imaging history and the LNT model. Similarly, Furukawa, Misumi, Cologne, and Cullings, 2016 proposed a Bayesian semiparametric model that uses LNT, threshold, hormesis, and hypersensitivity (Fig. C.1) extrapolation models to estimate cancer risk. Environmental pollution models (Vitolo et al., 2018) the impact of air or water pollutants (e.g., ozone, nitrogen dioxide, or smog) on mortality rates for a geographic area based on topography and exposure levels (Jerrett et al., 2005). Pharmacokinetics models use differential equations to model a drug's impact on patients' health based on the drug dose given, patient characteristics, and the exposure time(s) (Donnet & Samson, 2013). X. Liu et al., 2020 proposed a pharmacokinetics model based on the LSTM that successfully captured the temporal effects of a simulated drug. An important component of pharmacokinetics models (M. Liu et al., 2011) is that they model how the drug concentrations in the body change over time depending on the drug's properties and the patient's metabolism.

Based on the review of existing risk assessment models, current low-dose radiation risk models have three important limitations (National Research Council, 2006). First, the LNT model can only be used if the patient population has similar properties to the target population used for the model. Otherwise, population conversion factors must be applied which increases the uncertainty of the model. Second, the linear effect of radiation risk from high to low levels remains experimentally unchallenged. Alternative extrapolations models such as threshold, hypersensitivity, and hormesis (Fig. C.1) have also been proposed (National Research Council, 2006). As a result, the National Research Council of the National Academies recommends against using the LNT model to predict cancer risks for patients at low doses (National Research Council, 2006). Third, patients' medical histories are not considered when assessing risk from imaging exposure.

To address the limitations in existing imaging exposure risk models, the proposed risk assessment model needs to attend to the following requirements:

Requirement 1 (R1): Estimate patients' exposure from medical imaging across different anatomical regions and modalities. Previous studies (Berrington de Gonzalez et al., 2009; Mathews et al., 2013) estimated exposure from medical imaging using mean values from the literature.

Requirement 2 (R2): Analyze the duration between imaging events and the frequency of imaging. Currently, existing low-dose radiation risk models estimate patient risk from each scan independently and sum the risk together to estimate total risk (Gonzalez et al., 2013; Furukawa et al., 2016). The complex and non-linear interactions between exposure events are not analyzed.

Requirement 3 (R3): Analyze heterogeneous concepts in health records (e.g., diagnostic and exposure data). In existing assessment models (Gonzalez et al., 2013; Furukawa et al., 2016), patients' risk from medical imaging are based on exposure amount, age, and sex. Patients' medical histories are not used to estimate risk.

Requirement 4 (R4): Analyze medical histories with different start and end times, number of visits, and cover different time periods.

Requirement 5 (R5): Consider different exposure decay rates (e.g., half-life) because the long term effect of exposure from medical imaging remains an open research question. As a result, the model needs to provide health providers and patients risk calculations based on various exposure decay rates.

Requirement 6 (R6): Calculate a patient's cancer risk with and without the proposed scan based on their medical and imaging history.

Requirements 1 and 2 were investigated in this thesis as the the first steps towards developing the medical imaging risk assessment DSS. To address R1 (Fig. 1.1b), I investigated the challenges of using imputation models to estimate missing data in health records (Chapter 3 and 4). To address R2 (Fig. 1.1c), I investigated the challenges of analyzing temporal health records using deep learning (Chapter 5). Based on the findings, a proof-of-concept transformer diagnostic model that addresses R1 - R4 is proposed and evaluated in Chapter 5 (Fig. 1.1d). Analyzing patients' cancer risks under different decay rates (R5), calculating patients' cancer risk from medical imaging (R6), and evaluating the proposed cancer risk model's performance compared to models from the literature was outside the scope of this thesis and remains for future work.

2.2 Imputation Models and Evaluation

The first objective in this thesis was to evaluate imputation models to estimate patients' exposure from medical imaging (Fig. 1.1b). To achieve this objective, the evaluation methodology and metrics used to assess imputation models was investigated. In this section, imputation models (Sect. 2.2.1) as well as the evaluation methodology and metrics (Sect. 2.2.2) are reviewed.

2.2.1 Imputation Models

Consider a matrix D containing data for r instances described by f features. The objective in inferential statistics is to estimate population parameters Q such as mean, variance (σ), and regression coefficients (θ) by calculating statistics $\hat{Q} = (\hat{\mu}, \hat{\sigma}, \hat{\theta})$ from D. D can also be used to train deep learning models. However, D may contain observed ($D_{(1)}$) and missing ($D_{(0)}$) data. Together, $D = (D_{(1)}, D_{(0)})$ is the matrix with complete data. The response matrix $R_{r,f}$ shows the locations of observed ($R_{r,f} = 1$) and missing values ($R_{r,f} = 0$). The missing data pattern of R (Little & Rubin, 2019) can be described as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data are MCAR when the probability of data being missing depends only on the overall probability of data being missing (ψ). Data are MAR when the probability of missing data depends on ψ and $D_{(1)}$. Data are MNAR when the probability of missing data depends on ψ, $D_{(1)}$, and $D_{(0)}$.

We can estimate missing data by drawing synthetic observations from the posterior distribution of the missing data, given the observed data and the process that generated the missing data. Formally, the posterior distribution is denoted as $P(D_{(0)}|D_{(1)}, R)$. Rubin, 1976 demonstrated that R and the process that generated

the missing data are ignorable when data are MCAR or MAR. In these cases, the distribution of D is assumed to be the same in $D_{(0)}$ and $D_{(1)}$ (Rubin, 1976). As a result, we can model the posterior distribution using the observed data and then use this model to create imputations for the missing data ($P(D_{(0)}|D_{(1)}, R) = P(D_{(0)}|D_{(1)})$). Note we need to include R and the process that generated the missing data in the model of the posterior distribution ($P(D_{(0)}|D_{(1)}, R)$) when data are MNAR.

Mean, statistical, and deep learning imputation models have been proposed to estimate missing data. Mean imputation estimates missing data using the mean (μ) values from the medical literature or complete samples in the dataset. While mean imputation maintains the sample size, it reduces the variability in the dataset (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012). On the other hand, statistical imputation models such as logistic regression, decision trees, sequential regression (Van Buuren & Groothuis-Oudshoorn, 2010), and predictive mean matching (PMM; Heymans and Eekhout, 2019; Rubin, 1986 impute missing data based on the remaining values in the dataset. PMM constructs separate multiple Bayesian linear regression models for each feature based on the complete instances in a dataset. The differences between each imputed estimate and all observed values are then calculated (Rubin, 1986). The final imputed value is randomly drawn from the five complete cases with the smallest differences to the imputed estimate. A benefit of PMM is that the imputation model constructs plausible estimates by replacing the imputed data with the closest values from the real data. However, PMM requires complete instances which limits the size of the training set when multiple features have missing data.

Imputation models based on deep learning such as denoising (DAE; Lall and Robinson, 2021), variational autoencoders (VAE; Nazabal, Olmos, Ghahramani, and

Valera, 2020), and generative adversarial nets (GAN; Yoon, Jordon, and van der Schaar, 2018) have been proposed. Multiple Imputation with Denoising Autoencoders (MIDAS; Lall and Robinson, 2021) is a denoising autoencoder that models the posterior distribution even when data is missing in multiple features (Lall & Robinson, 2021; Nazabal et al., 2020). MIDAS consists of an encoder to learn to code representation of the input in the latent space and a decoder that reconstructs the original input from the latent code. During training, missing data is introduced by dropping random inputs. In MIDAS, the training objective is to minimize the model's likelihood function or reconstruction error (Sinha, Pandey, & Pattnaik, 2018). The missing data is treated as noise that MIDAS removes (Lall & Robinson, 2021). Generative imputation models generate new instances from the posterior distribution of D that are closest to the missing data (Borji, 2019). Generative Adversarial Imputation Nets (GAIN; Yoon et al., 2018) is a deep learning imputation model consisting of a generator, discriminator, and hint generator. The generator is an autoencoder that learns to implicitly model the data distribution while the discriminator estimates the probability that a sample came from the data distribution. The discriminator has an output vector of length f (one per feature). The generator and discriminator are trained using an adversarial process. During training, the generator learns to improve the imputed values while the discriminator learns to better identify imputed instances. A hint generator provides the discriminator partial information on the original sample to focus the model's attention on certain features. As a result, the generator learns to generate features according to the posterior distribution to fool the discriminator. The training objective of generative models is to minimize the distance between the generated and original data distributions (Borji, 2019).

Deep learning imputation has three advantages over statistical imputation models. First, deep learning imputation models can model the distribution of a dataset without assumptions on the underlying data (e.g., distribution function). Second, missing data across multiple features can be estimated using a single imputation model. Third, deep learning models can capture the latent structure of complex high-dimensional data (e.g., the correlation between demographics, medication history, and clinical outcomes in EHR (Pham, Tran, Phung, & Venkatesh, 2017)).

Missing data can be estimated using single or multiple imputation (Rubin, 1988). Multiple imputation captures the uncertainty of the imputation model by performing $c > 1$ independent draws from the posterior distribution $P(D_{(0)}|D_{(1)})$ to generate c complete datasets. Each c imputed dataset is then analyzed and the average performance over all c datasets is calculated. For example, PMM generates c Bayesian coefficients for the regression model. MIDAS subsamples thinned networks from a trained model using dropout (Lall & Robinson, 2021). GAIN draws multiple synthetic examples from the estimated distribution (Yoon et al., 2018). Multiple imputation has been shown to have improved confidence intervals and p-values compared to single imputation (Van Buuren & Groothuis-Oudshoorn, 2010).

2.2.2 Evaluation of Imputation Models

No imputation model (mean, statistical, or deep learning) is ideal for imputing missing data in all datasets. As a result, the performance of the candidate imputation models on MIIDD must be evaluated. The evaluation methodology commonly used in the literature to assess the quality of an imputation model is as follows (MIT Critical Data, 2016):

1. Select the data subset with no missing values.
2. Introduce increasing rates of missing data (e.g., 2%-80%).
3. Estimate the missing data using imputation models.
4. Assess the imputation models using an evaluation metric.
5. Repeat steps 1 - 4 multiple times (e.g., five times).
6. Calculate and plot the average evaluation metric versus the rate of missing data.

Existing deep learning imputation models (Lall & Robinson, 2021; Yoon et al., 2018; Nazabal et al., 2020) were assessed using the above methodology.

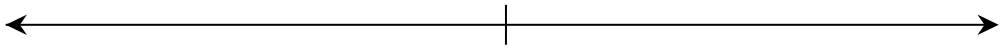
A commonly used metric to evaluate the quality of a deep learning imputation model (Step 4) is RMSE, which measures the difference between the imputed values and their corresponding actual values (Lall & Robinson, 2021; Nazabal et al., 2020; Yoon et al., 2018). Existing deep learning imputation models (MIDAS (Lall & Robinson, 2021), GAIN (Yoon et al., 2018), and VAE (Nazabal et al., 2020)) were assessed using RMSE. The studies showed the deep learning imputation models had competitive RMSE performance compared to statistical imputation models. However, the goal of imputation should not be achieving the best prediction accuracy by imputing the missing data, as in deep learning (Lall & Robinson, 2021; Nguyen et al., 2017), rather the goal is to ensure the imputed data meets the underlying properties of the data (e.g., data variability and distribution) (van Buuren, 2018). In addition, deep learning imputation models can impute missing data in multiple features at once. As a result, Lall and Robinson, 2021 and Yoon et al., 2018 assessed their deep learning imputation models' aggregated performances across all features with missing data. However, there are scenarios where specific features need to be imputed. For example, I needed to impute a target feature (low-dose radiation exposure) to develop the

risk model as the remaining features in MIIDD are complete. The deep learning imputation model's performance for target features has not been investigated. Finally, the deep learning imputation model's performance using qualitative and quantitative metrics has not been studied. In this thesis, two deep learning imputation models (DAE and GAN) are compared using qualitative, predictive accuracy, and statistical distance metrics on two tabular datasets.

Previous studies (Nguyen et al., 2017; Borji, 2019) have reviewed the performance metrics used to evaluate statistical and deep learning models. Borji, 2019 reviewed evaluation metrics to train and evaluate GANs for image generation. Borji demonstrated how qualitative and quantitative evaluation metrics assessed various aspects of the deep learning model's image generation. Similarly, Nguyen et al., 2017 reviewed evaluation metrics (qualitative, predictive accuracy metrics, and PCC) used to assess imputation models. Both studies recommended using different metrics to assess various properties of the imputation model's performance. However, deep learning imputation models have been evaluated using RMSE, a predictive accuracy metric.

2.3 Transformers for Health Record Analysis

The second objective in this thesis was to investigate using deep learning to analyze the elapsed time between medical events (Fig. 1.1c). In this section, the health records representations (Sect. 2.3.1) and transformer models (Sect. 2.3.2) used in existing predictive diagnostic models are reviewed.



- Time series
- Binning/Sequences
- Bag-of-words
- Window slicing
- Aggregation

Figure 2.2: Encoding temporal health records

2.3.1 Encoding Health Records

Electronic health records are tabular datasets that contain irregular observations of a patients' health. Previous studies have investigated encoding medical histories using: 1) Time series (Wiens, Horvitz, & Guttag, 2012; Hameed & Kleinberg, 2020), 2) Binning/Sequences (Xiao, Choi, & Sun, 2018), 3) Bag-of-words (Schafer, 2016), 4) Window slicing (Dietterich, 2002), and 5) Aggregation (M. Huang, Zolnoori, Shah, & Yao, 2018; Boursalie, Samavi, & Doyle, 2018) representations (Fig. 2.2). First, time series models encode each time step as an input feature, but EHR are modeled as regular time series with missing data. Second, binning/sequences represent medical records as ordered diagnostic or visit-level sequences of fixed duration (Choi, Xiao, Stewart, & Sun, 2018). Third, bag-of-words use a sliding window to generate an order invariant histogram of observations over patients' medical histories. Fourth, window slicing splits each patient's medical history into separate windows that are classified separately, and a majority voting approach is used to predict a patient's diagnosis. Finally, aggregation summarizes a patient's medical history into one vector which removes the temporal and sequential order of diagnoses. The elapsed time between medical visits is not included in these health record encodings.

2.3.2 Encoder-only Transformers

Analyzing irregular temporally-embedded datasets such as health records using machine learning models is a challenge. Early diagnostic models used the support vector machine (SVM; Boursalie, Samavi, and Doyle, 2015; Boursalie et al., 2018) and convolutional neural network (CNN; Fawaz, Forestier, Weber, Idoumghar, and Muller, 2018) to analyze aggregated medical histories which removed the temporal and sequential characteristics from health records. On the other hand, sequential deep learning models such as RNN (Rumelhart et al., 1985), LSTM (Hochreiter & Schmidhuber, 1997; Shickel et al., 2018), and transformers (Rasmy et al., 2021) have been used to analyze medical histories encoded as sequences (Pham, Tran, Phung, & Venkatesh, 2016; Devlin, Chang, Lee, & Toutanova, 2019). Transformers have three advantages over the RNN and LSTM (Vaswani et al., 2017). First, transformers access sequence elements directly rather than through the recursive memory. Second, transformers analyze the elements in each sequence at once, which enables parallelization to reduce training time. Third, transformer models can be trained on larger datasets.

There is growing interest in using transformers to analyze medical histories encoded as sequences. Consider a sequence $\{x_i\}$ where $i = \{0...j\}$ and j is the last element in the sequence. Each element i is dependent on the previous $i-1$ terms in the sequence. The encoder-only transformer (Vaswani et al., 2017), such as BERT (Bidirectional encoder representations from transformers; Devlin et al., 2019), is defined as:

$$y_j = f_{c1}(f_e(\{x_0, ..., x_{j-1}\})) \tag{2.1}$$
$$\{y_0, ..., y_{j-1}\} = f_{c2}(f_e(\{x_0, ..., x_{j-1}\}))$$

where y_j is the final element in the sequence predicted using the previous $j-1$ terms. During training, the encoder-only transformer also predicts a subset of randomly masked elements $\{y_0, ..., y_{j-1}\}$ using the remaining $\{x_0, ..., x_{j-1}\}$ terms to generalize the model, which is known as a masked language modeling (MLM; Devlin et al., 2019) The transformer maps the available input elements to a continuous representation $z = \{z_0, ..., z_{j-1}\}$ using an encoder layer. Each element in z is a weighted sum of the input to the encoder layer. The encoder-only transformer model is bi-directional so each attention head can attend to all positions in the sequence. The amount of weight (attention) applied to each input element is learned by the model during training. Examining the attention heads provides researchers a mechanism to investigate the sequence patterns learned by the model. Interested readers are referred to Vaswani et al., 2017 for more information on the attention mechanism. Encoder layers are stacked (f_e) to construct a high-level representations of the input sequence. The first element in the continuous representation of the final encoder layer (z_0) is fed to a feed-forward classification layer (f_{c1}) to predict the final element (y_n) in the sequence. In addition, the continuous representation of the final encoder layer for each masked elements $\{z_0, ..., z_{j-1}\}$ is fed through a separate classification layer f_{c2} to predict the masked input elements. The combined classification loss from f_{c1} and f_{c2} is used to update the model using back-propagation. The transformer has demonstrated improved performance for NLP (Vaswani et al., 2017) and diagnostic prediction (Y. Li et al., 2020) compared to the RNN and LSTM. Y. Li et al., 2020 and Rasmy et al., 2021 used BERT-based models to predict patients' final diagnosis based on their medical histories. However, transformers lack a mechanism to analyze the elapsed times between visits.

2.4 Summary

In this chapter, risk assessment models, imputation models and their evaluation metrics, and transformers for health record analysis were reviewed. Based on this review, three important challenges were identified that became the focus of the thesis:

1. **Impact of mean imputation on a deep learning model's performance has not been investigated (Fig. 1.1b):** This gap is the focus of Chapter 3.

2. **Deep learning-based imputation models (Lall & Robinson, 2021; Nazabal et al., 2020; Yoon et al., 2018) have been evaluated using RMSE, a predictive accuracy metric (Fig. 1.1b):** The deep learning-based imputation model's performance capturing the underlying properties of the dataset (e.g., mean and distribution) using evaluation metrics from the statistical literature has not been investigated. This gap is the focus of Chapter 4.

3. **Elapsed time between medical visits is not considered when transformers are used to analyze health records (Fig. 1.1c-d):** Instead, existing transformer-based diagnostic models in the literature (Y. Li et al., 2020; Rasmy et al., 2021) analyze the sequential order of medical diagnoses to learn disease patterns. This gap is the focus of Chapter 5.

Chapter 3

Mean Imputation in Deep Learning

An important step in evaluating the DTTHRE model was imputing the missing data in MIIDD. The MIIDD has approximately 2.1 million imaging records. Due to technical and privacy challenges, a representative subset of 39,909 DICOM headers was accessed that contain the scanner data required to estimate patient exposure. As a result, the exposure for the remaining imaging records in MIIDD must be imputed. A common method to impute medical imaging exposure is using mean values from the literature (Berrington de Gonzalez et al., 2009; Mathews et al., 2013). However, the predictive power of mean values to impute patients' exposure needs to be investigated.

In this chapter, the performance of mean imputation to estimate missing exposure data in health records and the impact on a deep learning model is investigated (Fig. 1.1b). A comparative analysis is performed between two methods to estimate low-dose radiation exposure from computed tomography (CT) and x-ray (XR) scans: 1) mean values from the literature and 2) calculated dose estimates from imaging scans. I also used techniques from adversarial machine learning (Papernot et al., 2016)

to demonstrate how the difference between estimation methods impacts a proof-of-concept deep learning model. The results show moderate increases in the mean values compared to the literature across all imaging exam types. However, using mean values reported in the literature underestimated patients' total exposure from medical imaging over the study period. The results also demonstrate that the discrepancies between the estimation methods was sufficient to cause model misclassification.

This chapter is structured as follows. In Section 3.1, an overview of the target feature for imputation, effective dose (ED; ICRP, 2007), is presented. Next, a comparative analysis estimating exposure from medical imaging using dose means from the literature and estimates calculated from imaging scans is provided in Section 3.2. In Section 3.3, I used techniques from adversarial machine learning to investigate how the discrepancies between estimation methods can impact the performance of a proof-of-concept deep learning model. A summary of this chapter is provided in Section 3.4. The author brings to the reader's attention that this chapter was published in Boursalie et al., 2020b and reproduced with permission from the IEEE.

3.1 Effective Dose Estimation

Effective dose is a metric to estimate the uniform whole-body dose that has the same nominal radiation risk compared to the nonuniform exposure from medical imaging (ICRP, 2007). Effective dose enables the comparison of radiation exposure between anatomical regions and modalities. The effective dose ($ED_{P,x}$) for each scan x per patient P is calculated using Eq. 3.1 (ICRP, 2007):

$$ED_{P,x} = \frac{1}{2} \cdot \sum_{T} \left(w_T \cdot G_T(female) + w_T \cdot G_T(male) \right) \tag{3.1}$$

Table 3.1: Scanner parameters and equations used to estimate ED. ©2020 IEEE

Scanner Parameters	**Unit**	**DICOM Tag**	**CT (Toshiba)**	**CT (GE)**	**XR (All)**
Manufacturer	-	(0008,0070)	X	X	X
Model	-	(0008,1090)	X	X	X
Age	Year	(0010,1010)	X	X	X
Peak kVp	V	(0018,0060)	-	X	X
$CTDI_{vol}$	mGy	(0018,9345)	X	-	-
I	mA	(0018,1151)	-	X	X
τ	ms	(0018,1150)	-	X	X
SPF	mm	(0018,9311)	-	X	-
TCW	mm	(0018,9307)	-	X	-
DSD	mm	(0018,1110)	-	-	X
ED	mSv	-	Eq. 3.1-3.2	Eq. 3.1-3.3	Eq. 3.1 and 3.4

where w_T are weighting factors for each organ or tissue T from the International Commission on Radiological Protection (ICRP, 2007). G_T is the gender-specific equivalent dose to each organ or tissue. Note that the weighting factors used to estimate ED are averaged over age and gender (McCollough, Christner, & Kofler, 2010). As a result, ED is a generic risk estimate and not the risk to a specific patient.

Calculating G_T is modality and scanner-specific, as shown in Table 3.1. For CT, the National Cancer Institute Dosimetry System for CT (NCICT) v2.1 (C. Lee, Kim, Bolch, Moroz, & Les, 2015) was used to estimate G_T using the following equation:

$$G_T = \sum_R w_R \cdot \left(\sum_{u=SSL}^{u=SEL} DC_{CT}(organ, age, sex, kVp, u) \cdot CTDI_{vol} \right) \quad (3.2)$$

where SSL and SEL are the scan start and scan end locations, respectively. DC is a matrix of Monte Carlo calculations based on organ scanned, age, sex, peak x-ray tube voltage (kVp), and scan slice number u (C. Lee et al., 2015). w_R is a radiation

weighting factor proposed by ICRP, 2007. CTDI_{vol} is the volume CT dose index reported by modern scanners. CTDI_{vol} can also be derived for older scanners using Eq. 3.3 (C. Lee et al., 2015):

$$CTDI_{vol} = \frac{nCTDI_w(manufacturer, model, spectrum)}{SPF} \cdot (\frac{I \cdot \tau}{100}) \cdot k_{OB} \quad (3.3)$$

where $nCTDI_w$ is the volume CT dose index selected from E. Lee, Lamart, Little, and Lee, 2014 based on manufacturer, model, and spectrum (combination of tube potentials and filtrations for a particular CT scan). SPF is spiral pitch factor, I is tube current, τ is exposure time, and k_{OB} is the overbeaming correction factor defined by Reiser, Becker, Nikolaou, and Glazer, 2008. Interested readers are referred to Huda and Mettler, 2011 for more information on the CT imaging scanner parameters used to estimate ED.

For XR, the CalDose_X v5.0 tool (Kramer, Khoury, & Vieira, 2010) was used to estimate G_T using the following equation:

$$G_T = \sum_R w_R \cdot DC_{XR}(exam, position, projection, DSD, I \cdot \tau, kVp, age, sex) \quad (3.4)$$

where DC is a matrix of Monte Carlo calculations based on imaging exam, patient position (standing or supine), image projection, distance source to detector (DSD), age, and sex.

Each imaging scan in this study was mapped to one of the exam types defined in the European Commission Report N154 (European Commission, 2008) for CT (head, neck, chest, abdomen, pelvis, and trunk) and XR (chest, cervical/thoracic/lumbar spine, abdomen, and pelvis). Each imaging study contains multiple scans representing

separate exposure events (e.g., a study has three scans). As a result, I calculated the effective dose for each scan in a study separately. Interpolated imaging scans (e.g, reconstructed, sagittal, and coronal) do not represent exposure events and were excluded from the study. Imaging scans with missing dosage information (Table 3.1) were also excluded. The distribution and mean effective dose for each exam type e ($\mu_{HAM,e}$) was calculated and compared to the EU mean dose ($\mu_{EU,e}$). Outliers beyond three standard deviations for each exam type were removed from the study.

In addition to calculating mean doses per exam type, each patient's total effective dose over the ten year study period was calculated. Two methods are used to estimate total effective dose: 1) effective dose means from the EU survey ($C_{P,EU}$; European Commission, 2015) and 2) effective dose calculated from imaging records ($C_{P,HAM}$) using Eq. 3.5 and 3.6, respectively. $N_{P,e}$ is the total number of scans per exam type e for patient P and N_P is the total number of scans per patient. The percentage change per person (PC_P) is calculated using Eq. 3.7.

$$C_{P,EU} = \sum_{e=1}^{E} \mu_{EU,e} \cdot N_{P,e} \tag{3.5}$$

$$C_{P,HAM} = \sum_{x=1}^{N_P} ED_{P,x} \tag{3.6}$$

$$PC_P = \frac{C_{P,HAM} - C_{P,EU}}{C_{P,EU}} \cdot 100 \tag{3.7}$$

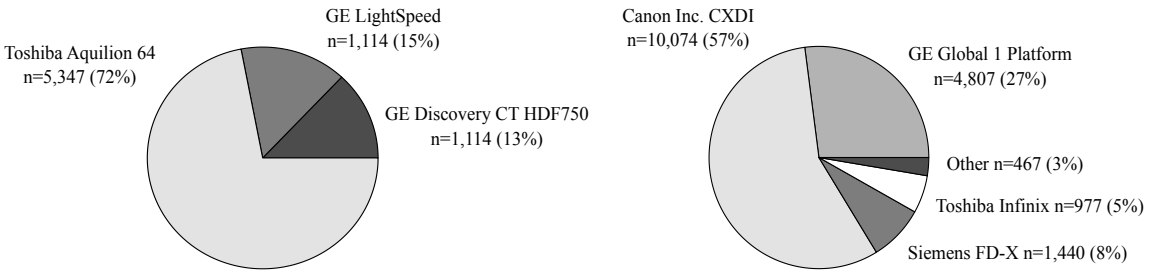


Figure 3.1: Distribution of CT (left) and XR (right) scans by model. ©2020 IEEE

3.2 Experimental Evaluation

In this section, the performance of mean imputation to estimate patients' ED exposure from individual medical scans and their total ED over the study period is evaluated.

3.2.1 Data Collection

A retrospective study was performed of all medical imaging scans from 2,000 patients who received at least one low-dose scan (e.g., CT and XR) from four hospitals in Hamilton, Ontario, Canada between May 2006 - 2016. The patients were a stratified random sample representative of the patient cohort in MIIDD in terms of age of first scan, sex, and body part scanned. In addition, the patients had above-average cumulative ED exposure. This study was approved by the University ethics board.

3.2.2 Evaluation of Mean Imputation

The representative study sample of 2,000 patients had 18,875 imaging studies (5,357 CT and 13,518 XR studies) with 39,909 imaging scans (7,427 CT and 32,482 XR scans) that resulted in low-dose radiation exposure. The breakdown of imaging scans by manufacturer and model are shown in Fig. 3.1. The majority of CT scans (72%)

were from Toshiba Aquiline 64 scanners and the ED was estimated using Eq. 3.1-3.2. The remaining CT scans (28%) were from GE scanners and Eq. 3.1-3.3 were used to estimate ED. The ED for XR scans were estimated using Eq. 3.1 and 3.4.

Figure 3.2 shows the estimated effective dose distributions, mean, and scan areas for each CT and XR exam type. Chest, abdomen, and trunk scans had similar ED distributions despite having different scan areas. Head CT scans had the lowest mean ED (3.14 mSv) and the smallest distribution (1-5 mSv). Chest, abdomen, and trunk CT examinations had the highest mean effective doses (12-17 mSv). Abdomen, thoracic and lumbar spine XR scans had the highest mean ED and similar distributions. Unlike CT, chest and cervical spine XR scans (including neck) had the lowest effective dose estimations. X-ray ED estimations were the result of all scan projections (e.g., anterior-posterior, posterior-anterior, left/right lateral, and posterior oblique).

The mean effective doses were then compared to the EU survey results. The estimated mean CT effective dose for head, abdomen, and trunk agreed with the mean doses in the EU survey. Similarly, the estimated mean for XR chest, spine (cervical, thoracic and lumbar), and abdomen had good agreement with the EU survey means. On the other hand, CT neck, chest, and XR pelvis estimates were higher than the EU survey. In fact, the CT chest and XR pelvis estimate means were around two times greater than the EU means. All EU effective dose means were within the distributions for each exam type. The EU survey did not include the raw data, so statistical comparisons such as t-test cannot be performed. The differences between the mean doses may result from different scanning equipment, imaging protocols, patient sizes, and training regiments across different institutions (Osei & Darko, 2013).

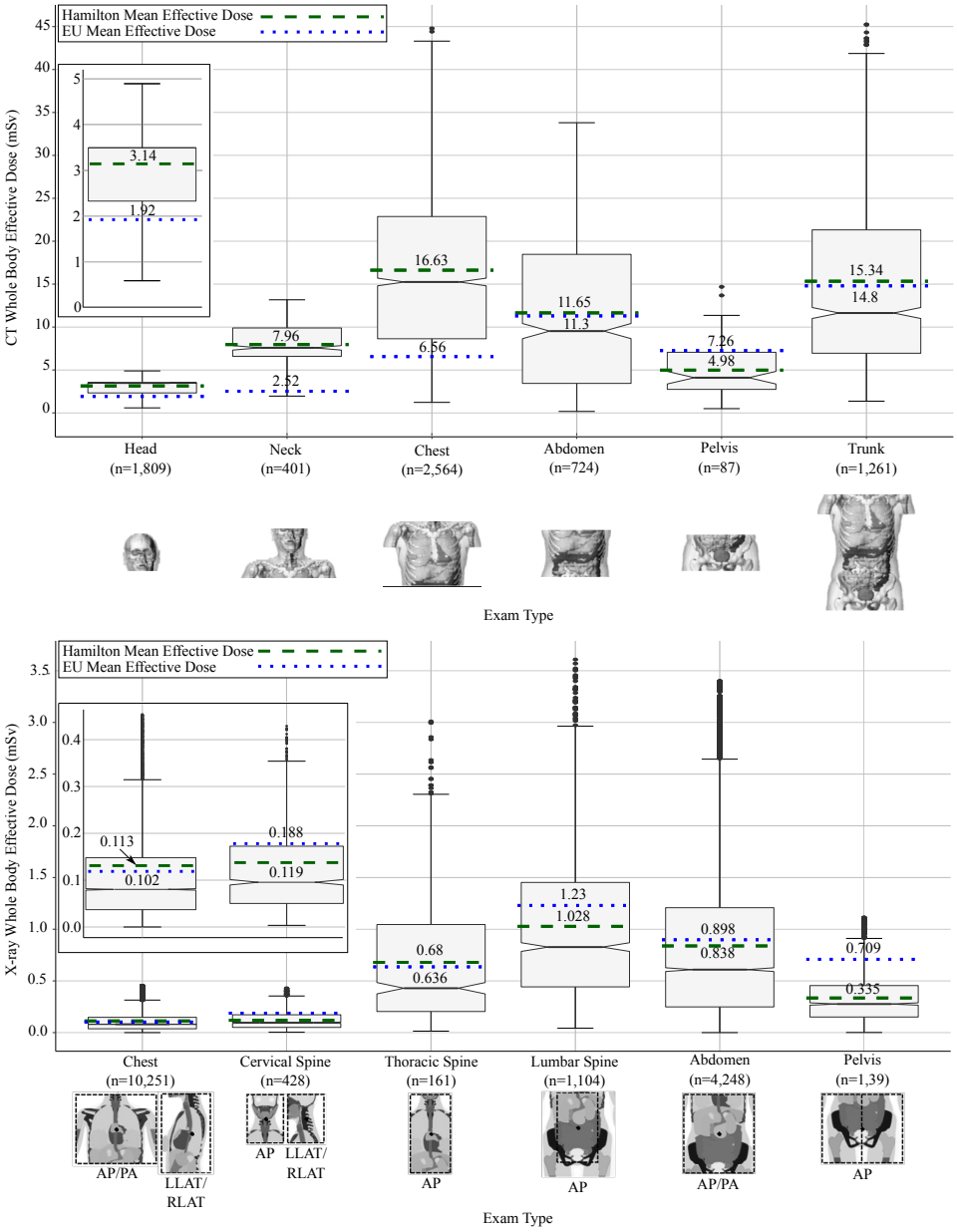


Figure 3.2: Boxplot of CT (top) and XR (bottom) ED per exam type. Lines denote the mean ED values from the literature (blue dashes) and this study (green dots). Lines within boxplot boxes denote medians; notch denotes 95% confidence interval of the median; bottom and top borders of boxes denote the 25^{th} and 75^{th} percentiles, respectively; vertical lines denote ranges (excluding outliers); and circles denote outliers. The CT and XR scan areas (modified from C. Lee, Kim, Bolch, Moroz, and Les, 2015 and Kramer, Khoury, and Vieira, 2010, respectively) are shown below each exam type. X-ray lumbar spine posterior oblique projections (left and right) are not shown.

Two design decisions in this study were identified that contributed to the discrepancy between estimation methods. First, the definitions of the start and stop positions impact the effective dose estimations because they determine which organs or tissues are within the scan area for the ED calculator. The scan area definition was a source of bias because scan definitions vary between institutions. For example, the NCICT calculator default scan ranges are defined from the protocols at the National Institute of Health Clinical Center in Bethesda, Maryland (C. Lee et al., 2015). The chest scan area covers the clavicles to the bottom of the liver (Fig. 3.2). If we define the chest scan area from the clavicle to the top of the liver, the ED estimation of the CT chest scan is reduced by 20%. Similarly, the x-ray field positions can be defined in CalDose_X. Second, the different scanner models impacted the NCICT effective dose calculations. For example, the effective dose for a male adult chest scan from Toshiba Aquilion 64 Slice CT scanners was 10.49 mSv. On the other hand, the effective dose estimation was 5.96 mSv from a Philips Brilliance 64 Slice CT scanner, decreasing the average dose estimation by 56%.

Two main sources of bias were identified in this study. First, the NCICT tool was used to estimate ED rather than the IMPACT calculator used in the EU dose survey (European Commission, 2015). To investigate the bias from the NCICT calculator, I compared the NCICT ED values with the estimations from the IMPACT calculator. Second, different DICOM attributes for the Toshiba and GE CT scanners were used to estimate ED. This bias was investigated by estimating ED using age, kVp, x-ray current, exposure time, SPF, and total collimation width (TCW) for all CT scans in this study. HHS's protocol values for SPF and TCW were used for Toshiba scans because the DICOM attributes are not recorded by the manufacturer.

3.2.3 Evaluation of Total Patient Exposure Estimation

Figures 3.3a-b and 3.3d-e show the estimated CT and XR total effective doses for each patient using the EU and HAM estimation techniques, respectively. The percentage difference between the EU and HAM cumulative ED estimations for the CT and XR scans are shown in Fig. 3.3c and 3.3f, respectively. Interestingly, there was a difference between the total ED from each estimation method despite similar mean dose values (Fig. 3.2). The total CT effective doses for 66% (807/1,223) of patients with a CT scan history were overestimated using mean dose values from the literature while 34% (416/1,223) of patients were underestimated. On the other hand, the total x-ray ED for 45% (798/1,775) and 55% (977/1,775) of patients with XR scan histories were over and underestimated, respectively, using mean dose values from the literature.

This study demonstrates the challenges of using mean ED values from the literature to estimate patient exposure from medical imaging. Mean effective doses were a reasonable estimate of a patient's average exposure from a single scan of a particular exam type. However, the difference between patients' total ED estimates from the mean values from the literature and calculated from medical images accumulated over time. For example, the number of scans per study is not captured in the health, imaging, insurance, or billing records used to estimate patients' total exposure (Eq. 3.5). As a result, researchers may assume that each study contains one imaging scan that results in patients' low-dose radiation exposure. The findings in this study show that CT and XR studies on average contained two imaging scans that results in patients' low-dose radiation exposure (CT: range [1:13], XR: range [1:24]) which impacts the total dose estimation. Ideally, a patient's total effective dose should be estimated

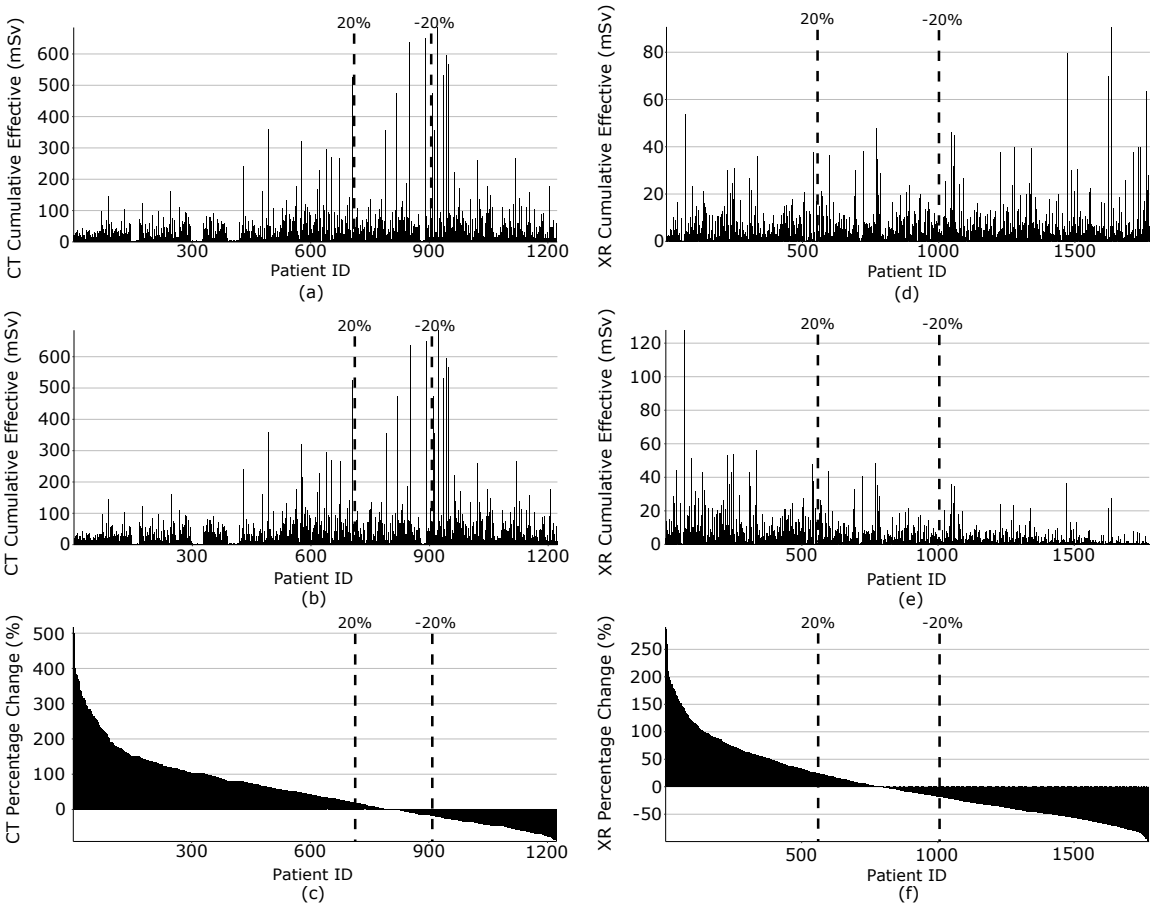


Figure 3.3: The cumulative CT and XR ED estimation for each patient P using literature values is shown in (a) and (d), respectively. Next, the cumulative CT and XR ED estimation for each patient from the ED tools is shown in (b) and (e), respectively. Then, the PC between each estimation method for CT and XR are shown in (c) and (f), respectively. Dotted lines show the $\pm 20\%$ PC

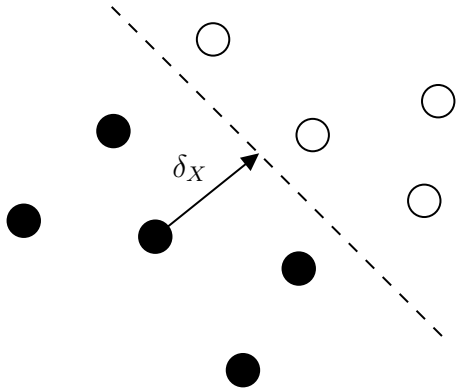


Figure 3.4: Adversarial attack where a perturbation δ_X results in misclassification

using each patient's DICOM headers and images. Patient-specific dose reconstruction (Tian, Yin, Man, & Samei, 2013; De Man, Wu, FitzGerald, Kalra, & Yin, 2015; Wu, Yin, & De Man, 2017) is a promising approach to more accurately estimate patients' ED exposures based on their body size and anatomy without computationally expensive Monte Carlo simulations. Estimating effective dose for a large study sample over a long period of time is challenging because it is difficult to collect the medical images, protocols, and machine parameters for all imaging procedures.

3.3 Effect of Mean Imputation in Deep Learning

In adversarial machine learning, an attacker changes the input feature vector ($\mathbf{X}$) by a perturbation (δ_X) to generate an adversarial sample ($\mathbf{X^*} = \mathbf{X} + \delta_X$) that results in a model misclassification (I. J. Goodfellow, Shlens, & Szegedy, 2015), as shown in Fig. 3.4. Previous studies demonstrated that deep learning models are sensitive to small input perturbations (I. Goodfellow, Shlens, & Szegedy, 2016). Interested readers are referred to Vorobeychik and Kantarcioglu, 2018 for more information on adversarial machine learning.

In this study, the perturbation represents the difference between the mean effective doses from the literature used for imputation and the calculated dose estimates from imaging scans. A proof-of-concept deep learning model was developed to investigate whether the impact of the two different imputation methods was similar to the impact of the input set being perturbed by an adversary. The deep learning model classified each patient's exposure history as above or below the mean EU CT cumulative exposure ($\mu_{C_{EU(CT)}}$=37.81 mSv). Imaging records from patients with a CT scan history (n=1,214) were used to train the model. The training set has six features ($f_{P,s}$) describing each patient's total CT exposure per exam e (head, neck, chest, abdomen, pelvis, and trunk) between 2006-2017. The CT ED exposure was estimated using the EU mean values from the literature ($f_{P,e} = \mu_{_{EU,e}} \cdot N_{P,e}$). The model's architecture had an input layer with six neurons, as well as a hidden and output layer with two neurons each. This model architecture was selected because each input feature can be perturbed independently to change the model's classification. The model was trained using k-fold (k=10) cross-validation (CV) with 20% of the training data used as the validation set. The L_0 Carlini-Wagner attack strategy (Carlini & Wagner, 2017) was used to calculate the perturbation required per feature for each patient ($\delta_{P,f}$) to generate the adversarial examples for each test fold. The adversarial examples were constrained to the distribution range of calculated ED for each body part scanned. For example, an adversarial example of a chest CT scan was limited between 1.5 and 43 mSv. The model's performance against an increasing maximum admissible perturbation (δ_{max}; Melis et al., 2017) was also evaluated.

Figure 3.5 shows the absolute total perturbation ($|\delta_X|$) required to result in a misclassification by the model. Figure 3.6 shows the impact of perturbations on model

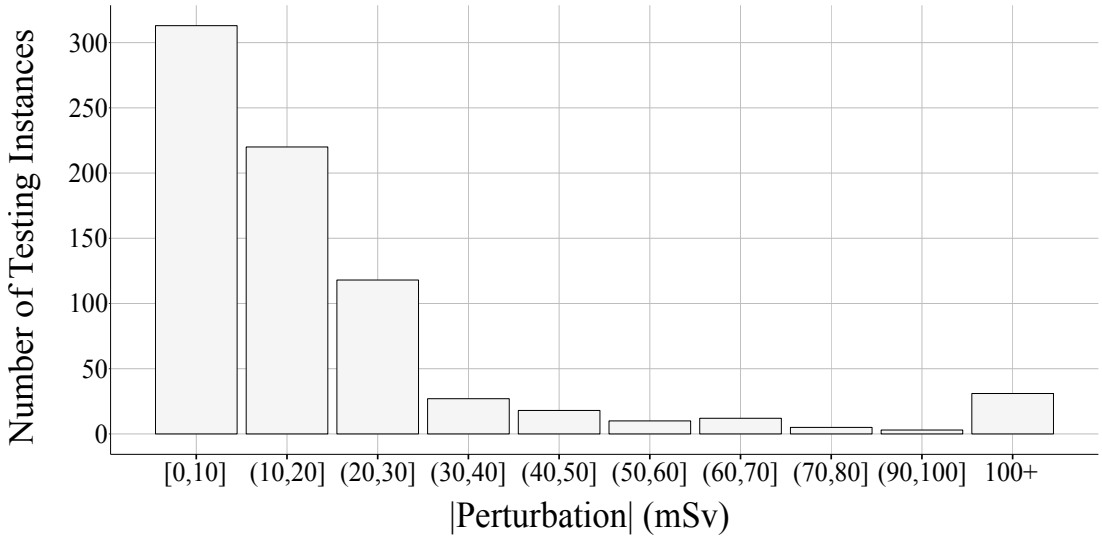


Figure 3.5: Histogram of the absolute perturbation required for patients with a CT scan history to result in a misclassification by the proof-of-concept deep learning model. ©2020 IEEE

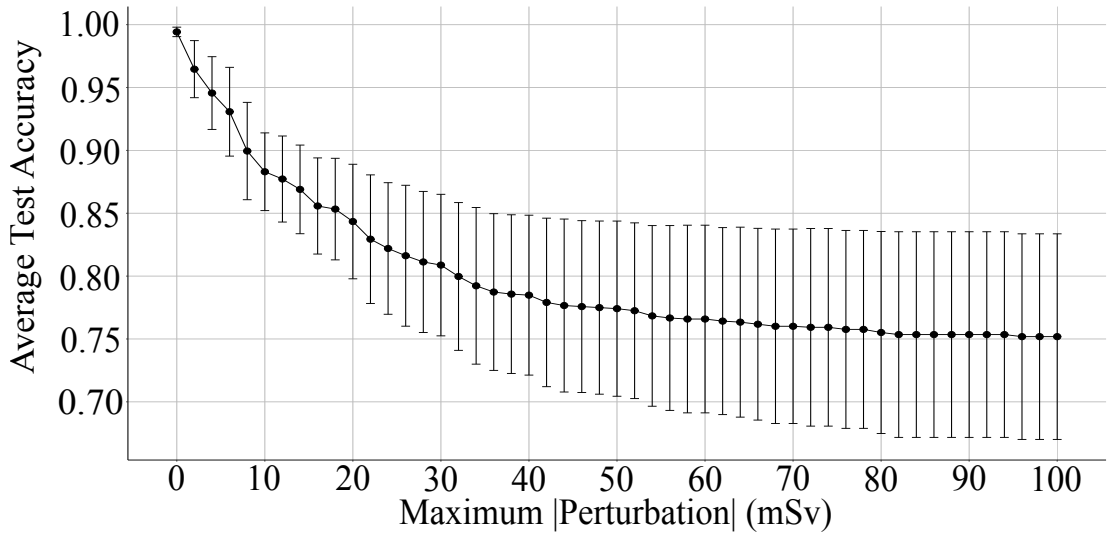


Figure 3.6: Average test accuracy ($k = 10$) of the proof-of-concept deep learning model against an increasing max input perturbation. The model's average test accuracy with no perturbation is shown at Max |Perturbation|= 0. ©2020 IEEE

performance. 57% (691/1,214) of patients with a CT scan history required a $|\delta_X| \leq$ 30 mSv over all scans in ten years to be misclassified. The perturbation can be distributed among any of the patient scans over ten years. A perturbation of less than 30 mSv over ten years decreased the models' average test accuracy by 20% (Fig. 3.6). The required perturbation was within the distribution ranges for each CT body part scanned (Fig. 3.2). Another focus in adversarial learning is examining the perturbations for instances closest to the decision boundary which are most vulnerable to misclassification. In this study, 26% (313/1,214) of patients with a CT scan history were borderline patients that required a $|\delta_X| \leq 10$ mSv over ten years to be misclassified. Such a perturbation can come from a single chest, abdomen, or trunk CT scan (Fig. 3.2). A perturbation of ≤ 10 mSv over ten years decreased the models' average test accuracy by 12% (Fig. 3.6). The results also show that the deep learning models' performance variability (error bars in Fig. 3.6) increased with δ_{max}. The increased variability reflects how the difference between estimation methods can accumulate based on the patient's number of scans and body part scanned. For example, head CT scans had a smaller distribution then chest scans (Fig. 3.2) so a patient's total head exposure will be more representative compared to their total chest exposure. The results demonstrate how small changes in the imputation method can impact a deep learning model's performance.

This study exhibits some limitations. Effective dose is an estimation of exposure that cannot be directly measured or validated (ICRP, 2007). As a result, we compared two techniques to estimate ED. In addition, the XR ED calculator does not leverage scanner manufacturer and model values which may contribute to their improved agreement with the literature mean dose estimations compared to CT.

3.4 Summary

Collecting complete medical histories for a large patient cohort over a long period is difficult due to technical, security, and privacy challenges. Imputation of missing data is an important step in developing diagnostic models such as DTTHRE. In this chapter, the mean imputation of ED was compared to estimations derived from medical images. Despite the good agreement between the mean values, using literature values underestimated patients' total ED. Techniques from adversarial machine learning were used to demonstrate the impact of imputation on a deep learning model.

I conclude with the following implications based on this study:

1. **Deep learning models are sensitive to the perturbations between imputed and actual values**: Perturbations from imputation models had similar effects on a deep learning model's performance as adversarial attacks.

2. **Discrepancies between the mean and actual values accumulated over a patient's medical history**: The accumulating discrepancies had a cascading negative impact on the performance of the deep learning model.

3. **It is important to consider the underlying feature properties (e.g., mean and distribution) when imputing datasets for analysis by deep learning models**: This is the focus of the next chapter.

Chapter 4

Evaluation Metrics for Deep Learning Imputation Models

In Chapter 3, I demonstrated that mean imputation underestimates patients' exposure from medical imaging. Another approach is to impute the exposure for the 2.1 million imaging records in MIIDD using a deep learning imputation model trained on the representative subset of 39,909 DICOM headers and diagnostic records. I was interested in using deep learning for imputation because the models make no assumptions about the underlying distribution of the data (Pham et al., 2017). Existing deep learning imputation models (Lall & Robinson, 2021; Nazabal et al., 2020; Yoon et al., 2018) were assessed using RMSE, which evaluates predictive performance.

In this chapter, the limitations of the evaluation metric RMSE for assessing the performance of deep learning-based imputation models is investigated (Fig. 1.1b). I also review and assess metrics from the statistical literature (qualitative, predictive accuracy, and statistical distance) to evaluate deep learning-based imputation models. A comparative analysis was conducted of two deep learning imputation models

(MIDAs and GAIN) and a regression model (PMM) using two datasets from different industry sectors: healthcare and financial. The results of the comparative analysis show that contrary to the commonly used RMSE metric, the statistical metric Jensen Shannon Distance (JSDist; Briet and Harremoes, 2009) best assessed the imputation models' performances. The regression model also ranked higher than deep learning when evaluated using the JSDist metric.

This chapter is organized as follows. In Section 4.1, a survey of the available evaluation metrics from the statistical literature to evaluate imputation models is described. Next, a comparative analysis of deep learning imputation models using the reviewed metrics is presented in Section 4.2. The limitations of using the evaluation metrics to assess deep learning imputation models are discussed in Section 4.3. A summary of this chapter is provided in Section 4.4. The author brings to the reader's attention that this chapter has been accepted for publication in Boursalie et al., 2021b and reproduced with permission from Springer.

4.1 Evaluation Metrics

In the statistical literature, the evaluation metrics can be qualitative (e.g., histogram, box, and density plots) or quantitative (e.g., predictive accuracy and statistical distance), as shown in Table 4.1. Predictive accuracy metrics measure the differences between the imputed values and their corresponding actual values. RMSE is a predictive accuracy metric and is defined in Eq. 4.1 where $\hat{b}$ and b are the imputed and actual values for r observations, respectively. Smaller RMSE indicates better agreement between the imputed and actual values.

Unlike predictive accuracy metrics, statistical distance metrics such as Cohen's

Table 4.1: Evaluation metrics summary. Reproduced with permission from Springer

Metric Type	**Metric**	**Description**	**Assumptions**	**Used**
Qualitative	Histogram	Graph of distributions		(Nguyen, Carlin, & Lee, 2017)
Predictive Accuracy	RMSE	Difference between the predicted and observed values	- Errors are unbiased and follow a normal distribution	(Lall & Robinson, 2021; Yoon, Jordon, & van der Schaar, 2018)
Statistical Distance	CDT	Magnitude of differences between 2+ groups	- Similar sizes - Similar SD	
	ϕ-divergence (KL & Jensen Shannon divergence, JSDist)	Dissimilarity between two probability distributions	- $X_r = 0$ means $Y_r = 0$	(Kingma & Welling, 2014; Nazabal, Olmos, Ghahramani, & Valera, 2020; Nowozin, Cseke, & Tomioka, 2016)

$$RMSE = \sqrt{\frac{\sum_{r=0}^{R}(\hat{b}_r - b_r)^2}{R}} \tag{4.1}$$

$$CDT = \frac{\mu_b - \mu_{\hat{b}}}{SD_{pq}} \text{ where } SD_{pq} = \sqrt{\frac{SD_b^2 + SD_{\hat{b}}^2}{2}} \tag{4.2}$$

$$JSDist = \sqrt{\frac{KL(p,s)}{2} + \frac{KL(q,s)}{2}} \text{ where } KL(a,b) = \sum_{i=0}^{a_{bins}} a_i \cdot log_2(\frac{a_i}{b_i}), s = \frac{p+q}{2} \tag{4.3}$$

Distance Test (CDT; Cohen, 1988) and ϕ-divergence measure the distance between the actual (p) and imputed (q) probability densities. The CDT is defined in Eq. 4.2 where μ and SD are the mean and standard deviations of the actual and imputed distributions. Distributions with small, medium, and large differences have a CDT ≤ 0.2, $0.2 < \text{CDT} \leq 0.5$, and $0.5 < \text{CDT} \leq 0.8$, respectively. ϕ-divergence metrics estimate the divergence between p and q using $D_\phi(p||q) = \int p(x)\phi(\frac{q(x)}{p(x)})dx$ where ϕ is a class of distance functions. Examples of ϕ are the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), KL approximate lower-bound estimator (Arbel, Zhou, & Gretton, 2021), and JSDist (Eq. 4.3). A JSDist = 0 indicates identical distributions while JSDist = 1 represents maximally different distributions.

4.2 Comparative Analysis

In this section, the comparative analysis of qualitative, predictive accuracy, and statistical distance metrics to assess two deep learning imputation models (MIDAS and GAIN) and a regression-based imputation model (PMM) on two tabular datasets is presented. MIDAS and GAIN represent non-generative and generative deep learning imputation models, respectively. PMM was selected as the benchmark model from the statistical literature.

4.2.1 Data Collection and Processing

The imputation models (PMM, MIDAS, and GAIN) were evaluated on two tabular datasets: 1) MIIDD and 2) Credit (Cr; Yeh and Lien, 2009). The MIIDD was collected in a retrospective study I performed of all medical scans from 1,200 patients who

Table 4.2: Continuous (C) and discrete (D) features from the MIIDD and Credit dataset. Reproduced with permission from Springer

Dataset	Instances	Years	Features	Target	Other Features
MIIDD	2,565 (4 hospitals, 1,200 patients)	May 2006 to May 2016	53 (31 C, 22 D)	Effective Dose (C)	1. Age of first scan (C), 2. Date (C), 3. Sex (D), 4-25. ICD-10 Diagnostic History (D), 26-47. Months since last diagnosis (C), 48-53. Number of CT and XR head, neck, chest, abdomen, pelvis, and trunk scans (C)
Credit (Yeh & Lien, 2009)	29,602 (1 bank, 29,602 clients)	May 2005 to Sept. 2005	30 (15 C, 15 D)	Age (C)	1. Sex (D), 2. Limit (C), 3-7. Education (D), 8-10. Married? (D), 11-16. Paid on time? (D), 17-20. Amount (C), 21-26. Amount paid (C), 27. Default next month? (D)

received at least one low-dose medical imaging scan (e.g., CT and XR) from four hospitals in Canada between May 2006 and May 2016. The patients were a stratified random sample representative of the target population in terms of age of first scan, sex, and body part scanned. The patients also had above-average cumulative ED exposure. All imaging scans were in the DICOM format. This study was approved by the Hamilton Integrated Research Ethics Board. Information on the Credit dataset is available at Yeh and Lien, 2009. The MIIDD and Credit datasets were selected for this study because they have continuous and discrete features with no missing data. The Credit dataset was also used to evaluate GAIN (Yoon et al., 2018).

Table 4.2 describes the characteristics of each dataset. Unlike previous studies (Lall & Robinson, 2021; Yoon et al., 2018), the imputation models' performances

on one target feature from each dataset were assessed to be consistent with the imputation evaluation methodology (Chapter 2.2.2). Effective dose ($f_{MIIDD,ED}$) and age ($f_{Cr,A}$) in the MIIDD and Credit dataset were selected for imputation because they are continuous features with non-normal distributions. There is also a relationship between the target and the remaining features to build the imputation model. For example, an important component in imputing temporal data such as EHR is the correlations between features across time (Rahman, Huang, Claassen, Heintzman, & Kleinberg, 2015). In this study, the ED exposure was related to the scan year as older scanners have higher exposure rates.

The target imputation features, effective dose ($f_{MIIDD,ED}$) and age ($f_{Cr,A}$), had non-normal distributions. As a result, the imputation models were evaluated using the original and quantile transform (QT; Pedregosa et al., 2011) features. The QT maps each quantile of the non-normal feature distribution to the corresponding quantile of the normal distribution (Beasley, Erickson, & Allison, 2009). Using QT, features with non-normal distributions can be analyzed using statistical tests (e.g., parametric) and machine learning models (e.g., Gaussian Naive Bayes) that require normal feature distributions. Machine learning models have also shown improved performance using QT features (L. Li, Song, & Yang, 2019).

4.2.2 Evaluation

The PMM, MIDAS, and GAIN imputation models were assessed using the evaluation methodology described in Chapter 2.2.2 to impute missing data in the MIIDD and Credit datasets. Increasing proportions of data MCAR (2%, 4%, 8%, 10%, 20%, 40%, and 80%) were introduced in ED (MIIDD) and age (Credit). Data MCAR was

selected to be consistent with previous studies (Lall & Robinson, 2021; Yoon et al., 2018). Missing data was imputed at each proportion using PMM, MIDAS, and GAIN. The mean results ($c = 5$) were taken from the multiple imputation models (MIDAS and PMM) to compare performance with GAIN. Previous studies (van Buuren, 2018) have demonstrated that the imputation results did not significantly change when $c > 5$. The evaluation was then repeated five times. Data was randomly removed each time. The imputation models were investigated using four evaluation metrics: 1) Histogram (benchmark), 2) RMSE, 3) CDT, and 4) JSDist. The evaluation metrics represent qualitative (histogram), predictive accuracy (RMSE), and statistical distance (CDT and JSDist). The qualitative performances (histogram) were plotted for each run. In addition, the average RMSE, CDT, and JSDist performance over the five runs were plotted. All experiments were conducted on a 64-bit Windows 7 laptop with a 2.8 GHz Intel Xeon CPU and 16 GB RAM. The default PMM (Van Buuren & Groothuis-Oudshoorn, 2010), MIDAS (Lall & Robinson, 2021), and GAIN (Yoon et al., 2018) models were implemented using their respective open-source codes.

Figure 4.1 shows a subset of the qualitative histogram results (Full results are shown in Fig. D.1 and D.2). $f_{MIIDD,ED}$ and $f_{Cr,A}$ had non-normal distributions. The No-QT-PMM, No-QT-MIDAS, and No-QT-GAIN models did not capture the distribution of $f_{MIIDD,ED}$ and $f_{Cr,A}$. The imputed values from the non-generative models (PMM and MIDAS) had more normal distributions centered on the average values of the target features. On the other hand, the generative model (GAIN) suffered from mode collapse (Srivastava, Valkov, Russell, Gutmann, & Sutton, 2017). Mode collapse occurs when the GAIN discriminator does not distinguish well between the actual and imputed data. As a result, the GAIN generator learns to fool the

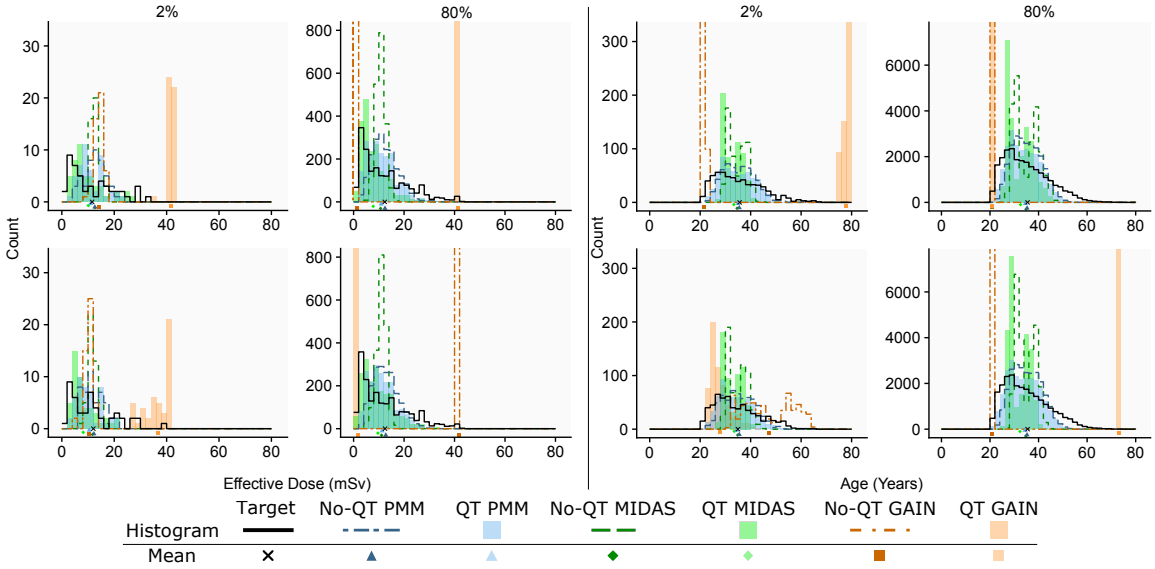


Figure 4.1: Histogram of $f_{MIIDD,ED}$ (left) and $f_{Cr,A}$ (right) imputation at 2% and 80% missing data over two runs (rows). Reproduced with permission from Springer. Full results are shown in Fig. D.1 and D.2

discriminator by generating modes of data that are not representative of the feature distribution. The results show that PMM and MIDAS had improved performance when $f_{MIIDD,ED}$ and $f_{Cr,A}$ were represented using a QT. For PMM and MIDAS, the imputation models better captured the distributions of $f_{MIIDD,ED}$ and $f_{Cr,A}$. The QT-GAIN models did not capture the mode or distribution of the data.

The imputation models' RMSE performances imputing $f_{MIIDD,ED}$ are shown in Fig. 4.2a. The No-QT-MIDAS model had the best RMSE performance across all missing data rates. Next, QT-MIDAS had the second-best RMSE performance. Then, the No-QT-PMM and QT-PMM models performed third best overall. The No-QT-GAIN model had a similar performance to the MIDAS and PMM models for 2%-40% missing rates. However, the No-QT-GAIN model's mode collapse was poorly detected using RMSE. The No-QT imputation models captured the $f_{MIIDD,ED}$ mean

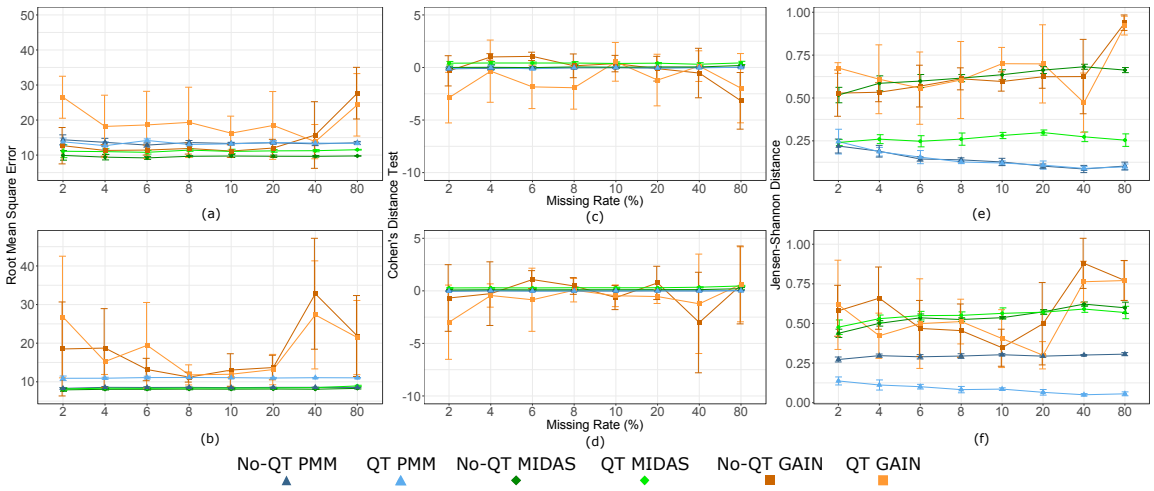


Figure 4.2: RMSE (a-b), CDT (c-d), and JSDist (e-f) evaluation results for $f_{MIIDD,ED}$ (top) and $f_{Cr,A}$ (bottom) at increasing missing data rates. Lines and error bars are average performance over five runs. Reproduced with permission from Springer

(Fig. 4.1), which minimized their RMSE. The QT-GAIN model captured the maximum $f_{MIIDD,ED}$ values, which resulted in the worst RMSE performance. The imputation models' RMSE performances were consistent on the Credit dataset (Fig. 4.2b) except for the No-QT-MIDAS, QT-MIDAS, and QT-PMM models, which had similar performances. Interestingly, the imputation models' RMSE performances did not agree with the qualitative results (Fig. 4.1). In addition, the improved distributional performances of the QT models were not captured using RMSE.

Figure 4.2c shows the CDT performances for the $f_{MIIDD,ED}$ imputation models. The No-QT MIDAS, No-QT-PMM, and QT-PMM models had the best CDT performances. Next, the QT-MIDAS had the second-best CDT performance. Then, the No-QT-GAIN and QT-GAIN models had the worst CDT performances. Similar to RMSE, the GAIN model's mode collapse was poorly detected using CDT. CDT compares the mean and standard deviations of the actual and imputed data (Eq. 4.2). As a result, the GAIN models had competitive and stable CDT performance despite not

capturing the $f_{MIIDD,ED}$ distribution. The $f_{Cr,A}$ imputation had similar CDT performance (Fig. 4.2d) across all models. On the other hand, the $f_{Cr,A}$ GAIN imputation models had the most unstable CDT results. Like RMSE, the CDT results for both datasets did not agree with the qualitative results (Fig. 4.1). Furthermore, the improved distributional performances of the QT imputation models were not captured using CDT.

The imputation models' JSDist performances imputing $f_{MIIDD,ED}$ are shown in Fig. 4.2e. The PMM models had the best JSDist performances for $f_{MIIDD,ED}$ imputation. Next, the QT-MIDAS had the second-best JSDist performance. Then, the No-QT-MIDAS, No-QT-GAIN, and QT-GAIN models had the worst JSDist performances. The No-QT-GAIN and QT-GAIN models also had the most unstable JSDist performance. Unlike RMSE and CDT, the JSDist was sensitive to detecting mode collapse in No-QT-GAIN and QT-GAIN. The GAIN model's reconstruction of the mean $f_{MIIDD,ED}$ value did not achieve competitive JSDist performance. On the Credit dataset (Fig. 4.2f), the QT-PMM model had the best JSDist performance while the No-QT-PMM model performed second-best. The No-QT-MIDAS, QT-MIDAS, No-QT-GAIN, and QT-GAIN models had the worst JSDist performances. Unlike the predictive accuracy metrics, the JSDist metrics for the $f_{MIIDD,ED}$ and $f_{Cr,A}$ imputation models agreed with the qualitative results. JSDist is a quantitative implementation of the qualitative comparison (Fig. 4.1), so the agreement between the evaluation metrics was understandable. The improved distributional performances of the QT imputation models were captured by the JSDist metric.

Table 4.3: Imputation models' ranked performances (1 best, 4 worst) based on the evaluation metrics. Reproduced with permission from Springer

	MIIDD (ED)						**Credit (Age)**					
	PMM		**MIDAS**		**GAIN**		**PMM**		**MIDAS**		**GAIN**	
	No-QT	**QT**	**No-QT**	**QT**	**No-QT**	**QT**	**No-QT**	**QT**	**No-QT**	**QT**	**No-QT**	**QT**
Histogram	**1**	**1**	**3**	**2**	**3**	**3**	**2**	**1**	**3**	**3**	**4**	**4**
RMSE	3	3	1	2	3	4	1	2	1	1	3	3
CDT	1	1	1	2	3	3	1	1	2	3	4	4
JSDist	1	1	3	2	3	3	2	1	3	3	4	4

4.3 Limitations of Performance Metrics

Table 4.3 ranks each imputation model's performance on each dataset based on the qualitative, predictive accuracy, and statistical distance metrics. The qualitative results were ranked based on a visual inspection of the mean and distribution reconstruction (Fig. 4.1) across all runs. Based on the histogram (benchmark) results, the QT-PMM would be selected for imputation in both datasets across all missing data rates. However, the predictive accuracy metrics did not agree with the qualitative and statistical distance results. The No-QT-MIDAS would be selected for the MIIDD and the No-QT-PMM, No-QT-MIDAS, or QT-MIDAS models would be selected for the Credit dataset based on RMSE. Using CDT, the No-QT-PMM, QT-PMM, or No-QT-MIDAS model would be selected for the MIIDD while the No-QT-PMM or QT-PMM would be selected for the Credit dataset. The JSDist ranking agreed with the histogram results.

The qualitative results (Fig. 4.1) provided an initial check of the imputation model's performance (Nguyen et al., 2017) and context to the quantitative metrics. For example, the qualitative results demonstrate that the poor performance of the

GAIN models was due to mode collapse. The qualitative results can also be reviewed by a domain expert. For example, a medical expert could review the qualitative results of the No-QT-PMM and QT-PMM models (Fig. 4.2e).

The results demonstrate that the predictive accuracy metrics evaluated the imputation model's ability to capture the mean of the target distributions. For example, RMSE directly compares the imputed estimates with the actual values rather than compare the distributions (Eq. 4.1). Similarly, CDT compares the means and standard deviations of the imputed and actual distributions (Eq. 4.2). Interestingly, the imputation models that generate more normal distributions (MIDAS and No-QT models) minimized their RMSE and CDT (Fig. 4.2a-d) without capturing the distribution of the target features. In fact, the PMM models that attempted to capture the distribution of the target features (Fig. 4.1) had low RMSE and CDT performances (Fig. 4.2a-d). In addition, the GAIN models demonstrate how competitive RMSE and CDT results (Fig. 4.2a-d) can be achieved by imputing a single value due to mode collapse. The results demonstrate how imputation models can achieve good predictive accuracy performance (Fig. 4.2a-b) without capturing the distributions of the features (Fig. 4.1).

The ϕ-divergence metric (JSDist) best assessed the imputation models' performances. Unlike predictive accuracy metrics, JSDist (Eq. 4.3) compares the target and imputed distributions rather than comparing imputed instances directly with their actual values. In this study, the target features had non-normal distributions. As a result, imputation models that generate more normal distributions (GAIN and No-QT) will diverge from the target feature distribution (Fig. 4.1). In addition, imputation models that capture the mean or mode of the data (Fig. 4.1) result in poor

JSDist (Fig. 4.2e-f) performances despite competitive RMSE and CDT (Fig. 4.2a-d) results. Divergence metrics have been used to evaluate generative deep learning models for image generation (Borji, 2019). The results demonstrate that ϕ-divergence metrics can also be used to evaluate deep learning-based imputation models.

The goal of imputation is not to predict the missing data in a dataset as in deep learning (van Buuren, 2018). Rather, the goal is to capture the underlying dataset properties (e.g., mean and distribution) that are hidden by missing data to prevent bias in the subsequent analysis. The qualitative, predictive accuracy and statistical distance metrics did not agree because they evaluated different qualities of the imputation model's performance. RMSE and CDT compared mean reconstruction while ϕ-divergence metrics examine the divergence between distributions. The results demonstrate that previous studies that evaluated deep learning imputation using predictive accuracy metrics (Lall & Robinson, 2021; Yoon et al., 2018) may not have captured the overall performance of their models.

The dataset properties may determine the suitability of the evaluation metrics to assess imputation models. For example, the statistical distance metric (JSDist) may better capture the difference between the target and imputed distributions for non-normal distributions compared to predictive accuracy metrics. Similar to statistical tests (Biau, Kere is, & Porcher, 2008), the choice of metric may depend on the dataset size. In this study, the metrics had similar performances (Fig. 4.2) between datasets with different sizes (Table 4.1).

Previous studies (Lall & Robinson, 2021; Yoon et al., 2018) evaluated their deep learning-based imputation model's aggregate performance across all features with simulated missing data. Unlike generative deep learning (Guan, Li, Yu, & Zhang,

2018), only a subset of the imputed values are used to replace the missing data for subsequent analysis. As a result, the aggregated performance may not represent the deep learning-based imputation model's performance on a single target feature.

This study exhibits the following limitations. First, the imputation models' performances on specific features (ED and age) were investigated. Second, data MAR and MNAR were not investigated. Third, the default model architectures were used. Improved model architecture and training could impact the performance ranking.

4.4 Summary

Deep learning imputations models are a promising approach to estimate missing data in health records. Imputation models introduce their own level of uncertainty that needs to be evaluated (Eekhout et al., 2012). In this chapter, I investigated the use of evaluation metrics from the statistical literature (qualitative, predictive accuracy, and statistical distance) to assess the performance of deep learning imputation models. I performed a comparative analysis of two deep learning imputation models (DAE and GAN) and a regression imputation model on two tabular datasets (healthcare and financial). The results show that the statistical distance metric (JSDist) best assessed the performances of the imputation models.

I conclude with the following implications for evaluating deep learning imputation models based on this study:

1. **Imputation is not prediction**: Deep learning imputation models need to be evaluated on their ability to capture underlying data properties instead of their predictive performance.

2. **Qualitative, predictive accuracy and, statistical distance metrics evaluated different properties of the imputation model's performance**: RMSE and CDT compared mean reconstruction while ϕ-divergence metrics examined the divergence between distributions.

3. **Selecting the evaluation metric for assessing deep learning-based imputation models depends on the dataset properties**: When selecting the evaluation metric, researchers should consider the dataset size, distribution of features, and proportion of missing data.

Chapter 5

Temporally-Embedded Deep Learning Model

An important characteristic of EHR is that they represent multiple episodes of a patient's point of care over time. In addition, the elapsed time between medical visits is irregular. One approach for analyzing the temporal patterns in medical histories is to encode each patient's medical history as a sequence of diagnostic codes (Y. Li et al., 2020; Rasmy et al., 2021). Transformers then analyze the sequence of diagnoses to learn disease patterns. However, transformers do not consider the elapsed time between medical visits when analyzing medical histories, which is an important factor when assessing patients' risk from medical imaging (National Research Council, 2006).

In this chapter, I propose and evaluate a decoder-only transformer model called Decoder Transformer for Temporally-Embedded Health Records Encoding (DTTHRE) that predicts the primary diagnosis for each visit using the patient's medical history, including the elapsed times between visits (Fig. 1.1c). The proposed model requires

an encoding mechanism that embeds the irregular time series data in medical histories. Instead of diagnostic-level encoding, I propose to encode medical histories as a sequence of medical events called Temporally-Embedded Health Records Encoding (THRE). A proof-of-concept DTTHRE model was then used to analyze the imputed MIIDD (Chapters 3 and 4) to investigate if embedding the time between visits impacts predictive performance (Fig. 1.1d). The performance of DTTHRE on the MIIDD was then compared to Med-BERT (Rasmy et al., 2021). DTTHRE successfully predicted patients' primary diagnosis in their final visit with improved predictive performance ($78.54 \pm 0.22\%$) compared to Med-BERT ($40.51 \pm 0.13\%$).

This chapter is organized as follows. In Section 5.1, the challenges of analyzing health records using transformers are reviewed. To address these challenges, the proposed DTTHRE model and THRE mechanism are described in Section 5.2. In Section 5.3, the evaluation of the proof-of-concept DTTHRE model is reported. A summary of this chapter is provided in Section 5.4. The author brings to the reader's attention that this chapter has been accepted for publication in Boursalie et al., 2021a and reproduced with permission from the IEEE.

5.1 Medical Records Characteristics and Decoder Transformers

In this section, the characteristics of health records and decoder-only transformers for diagnostic predictive models are reviewed.

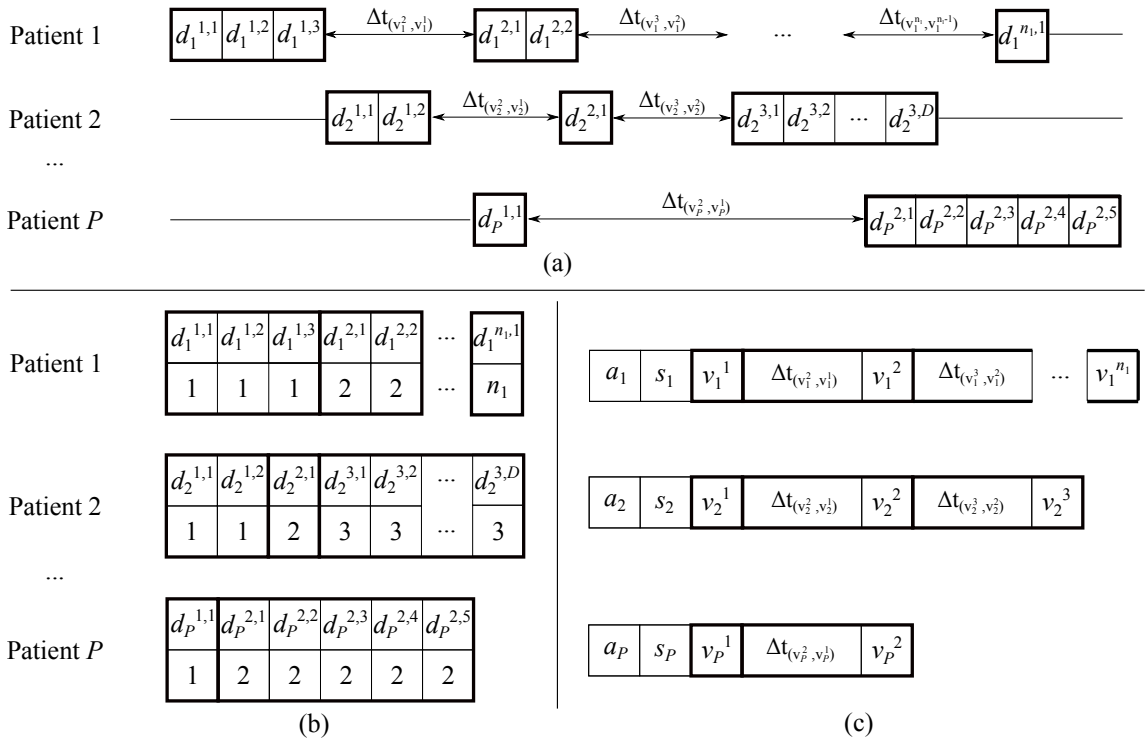


Figure 5.1: a) Raw, b) diagnostic-level, and c) THRE sequences of medical histories. Elements from the same visit are in the contiguous thick border blocks. ©2021 IEEE

5.1.1 Medical Records Characteristics

Electronic health records are tabular datasets that contain demographic, diagnostic, and treatment records for each patient. Medical histories contain irregular observations (Fig. 5.1a) as patients have different first and last visits, visit dates, number of visits, and number of diagnosis and treatments per visit. We define each $p = \{1...P\}$ patient's medical history as an irregular time series $H_p = \{v_{p,t_n}, n \in \mathbb{Z}\}$ with observation times $t_{n+1} > t_n$, $n \in \mathbb{Z}$ where v_{t_n} is the patient's n^{th} medical visit at time t_n. Each medical visit contains the following: 1) demographic data (age and sex), 2) diagnostic sequence of $d = \{1...D\}$ codes, and 3) medicine or treatment sequence of $m = \{1...M\}$ codes.

There is growing interest in encoding medical histories as sequences. Previous studies (Y. Li et al., 2020; Rasmy et al., 2021) represented patients' medical histories as sequences of diagnostic-level codes $h_p = \{d_p^{1,1}, ..., d_p^{n,D}\}$, as shown in Fig. 5.1b. In sequential modeling, disease progression is modeled using visit and diagnostic order ($o = (n, d)$) instead of time (t_n). The position embedding for each element is used to identify the corresponding visit. Representing medical histories as sequences enable researchers to use sequential models designed for NLP for disease prediction.

Y. Li et al., 2020 and Rasmy et al., 2021 demonstrated that encoding diagnostic codes as a sequence can be used for disease prediction without considering time. However, time of exposure is an important feature when assessing patients' risk from exposure due to medical imaging (National Research Council, 2006). In fact, encoding patients' irregular exposure history as a sequence would remove patterns that need to be modeled. In addition, encoding each diagnosis as a separate element (h_p) increases the search space as the model learns inter and intra-visit properties. There is a need to develop an encoding representation for diagnostic and exposure histories that incorporates time for analysis by transformer models.

5.1.2 Modeling Temporal Data Using Decoder Transformers

The transformer (Vaswani et al., 2017) is a deep learning sequential model that learns weight-adjusted representations for each element in a sequence. Previous studies (Y. Li et al., 2020) used encoder-only transformers to analyze patients' medical histories encoded as diagnostic sequences. The encoder-only transformer has bi-directional attention heads so a weight-adjusted representation for each element is constructed based on the remaining elements in the sequence.

The transformer is designed for NLP where the distance between subsequent elements is constant ($\lambda = C$). As a result, transformers cannot evaluate variable λ in time series data. For example, music is a regular time series that has varying λ_t between notes. The transformer assumes that $\lambda = C$ regardless of λ_t. One solution proposed by C. Huang et al., 2019 was to analyze piano music using a decoder-only transformer (Vaswani et al., 2017). A decoder-only transformer, such as the Generative pre-trained transformer (GPT-2; Radford et al., 2019), is defined as:

$$\{y_1, ..., y_i\} = f_{c3}(f_d(\{x_0, ..., x_{i-1}\})) \tag{5.1}$$

that predicts each element in the sequence ($\{y_1, ..., y_i\}$) using the previous $\{x_0, ..., x_{i-1}\}$ terms. Like the encoder-only transformer, the decoder-only transformer consists of stacked layers of multi-attention heads (f_d) that construct a continuous representation of the input sequence. However, the decoder masked multi-attention heads are uni-directional so each head pays attention only to the previous $i - 1$ elements in the sequence. In addition, each element of $\{z_0, ..., z_{i-1}\}$ in the final decoder layer is passed through the feed-forward classification layer (f_{c3}) to predict the next element $\{y_1, ..., y_i\}$ in the sequence. The combined classification loss from each predicted element is back-propagated through the model during training. The decoder model is autoregressive and generative. For example, GPT-2 (Vaswani et al., 2017; Radford et al., 2019) generates text by predicting each word in a sentence sequentially.

C. Huang et al., 2019 encodes the complex concepts in music (notes, velocity, and time shifts) as events in a musical sequence. Specifically, music was represented as MIDI (Musical Instrument Digital Interface) events (Oore, Simon, Dieleman, Eck, & Simonyan, 2020) that include NOTES_ON, NOTES_OFF, note VELOCITY, and

discretized TIME_SHIFTS in 10 ms increments. C. Huang et al., 2019 then trained a decoder-only transformer to generate music by predicting each musical event using the previous events in the sequence.

The method described in C. Huang et al., 2019 for analyzing time series music encoded as a sequence of events using a decoder-only transformer can be used to analyze medical histories if the following limitations are addressed:

Limitation 1 (L1): If we encode health records as an event sequence as in C. Huang et al., 2019, the decoder-only transformer would predict each medical event using previous events from the same visit which would bias the model. For example, consider a patient's visit where the first and third diagnoses have a hierarchical relationship. A decoder-only transformer could learn the multilevel relationships between diagnostic codes (Choi et al., 2018) that are not related to predicting disease. All events from the same visit must be analyzed by the decoder-only model at once to predict disease progression.

Limitation 2 (L2): The decoder-only transformer model in C. Huang et al., 2019 predicts each event in the music sequence. The goal of the diagnostic model is to predict the primary diagnosis for each patient's visit, not each event in the medical sequence (e.g., predict patient's age and sex by analyzing diagnoses).

Limitation 3 (L3): In NLP, the elements in the sequence are the same data type (words). Consequently, we sum the classification loss over all predicted elements to update the model using backpropagation. In contrast, elements in medical records contain multiple types of data such as age, sex, diagnoses, medication, and time between visits. The diagnostic model's classification loss must be evaluated on predicting the primary diagnosis for each visit, not the next sequence element.

5.2 DTTHRE Model

To address the gaps in the existing transformer diagnostic models, I propose: 1) a medical encoding representation called THRE and 2) a diagnostic model called DT-THRE. To address L1, the MIDI encoding described in Sect. 5.1.2 was extended for medical records. The health records irregular time series (H_p) were encoded as a sequence of medical events (Fig. 5.1c). There are four types of medical events in THRE: 1) AGE, 2) SEX, 3) VISIT, and 4) DELTA_TIME. AGE and SEX represent the patient's age at their first visit and sex, respectively. VISIT is the visit-level embedding of the diagnostic and medical/treatment observation sequences for each visit. Choi et al., 2018 demonstrated that visit-level embeddings improved predictive performance in deep learning models. DELTA_TIME represents the time between subsequent visits in monthly increments (Oore et al., 2020; C. Huang et al., 2019; Miotto, Li, Kidd, & Dudley, 2016). I also extended the decoder-only transformer (Sect. 5.1.2) to analyze medical events, as shown in Fig. 5.2. To address L1, the DTTHRE decoder layers were modified to analyze the event-pairs for each visit (DELTA_TIME and VISIT) rather than each element. To address L2 and L3, the feed-forward classification layer was modified to predict the next primary diagnosis (Table E.1) using the previous VISIT representation state instead of predicting the next i element using the last $i-1$ representation state.

Formally, THRE is expressed as the sequence:

$$THRE_p = \{a_p, s_p, v_p^1, \Delta t_{(v_p^2, v_p^1)}, v_p^2, ..., \Delta t_{(v_p^{n_p}, v_p^{n_p-1})}, v_p^{n_p}\}$$
$$\text{where } v_p^{n_p} = AE(\{d_p^{n_p}, m_p^{n_p}\}) \tag{5.2}$$

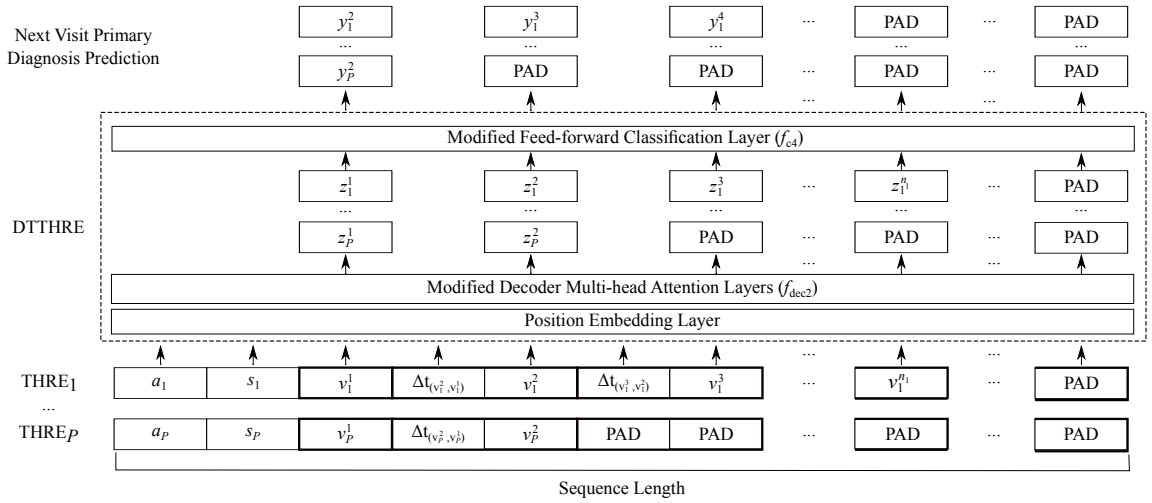


Figure 5.2: Proposed DTTHRE architecture. Elements from the same visit are in the contiguous thick border blocks. ©2021 IEEE

where a_p and s_p are each patient's AGE at first visit and SEX respectively, Δt is the time between subsequent visits (DELTA_TIME), and v_p is the visit-embedding (VISIT) representation of the diagnostic and medical/treatment observations for each visit. VISIT is generated by passing the concatenated diagnostic (d) and medical/treatment (m) observations for each visit through a trained auto-encoder (AE). The first visit for each patient is described by $\{a_p, s_p, v_p^1\}$. Subsequent patient visits are described using the event-pair $\{\Delta t_{(v_p^n, v_p^{n-1})}, v_p^n\}$.

DTTHRE is defined as:

$$\{l_4, l_6, ..., l_i\} = f_{c4}(f_{dec2}(\{x_0, ..., x_{i-1}\}) \tag{5.3}$$

where f_{dec2} is the modified uni-directional decoder layers, f_{c4} is the new feed-forward classification layer, and $\{l_4, l_6, ..., l_i\}$ are the labels for the target positions $\{4, 6..., i\}$ in x. In f_{dec2}, the three elements at the beginning of sequence $\{x_0, x_1, x_2\}$ (e.g., AGE,

SEX, and VISIT 1) are always available to the decoder attention heads. In addition, the masked self-attention heads were modified so pairs of elements are available to the attention heads instead of individual elements (Fig. 5.2). Instead of predicting the next element x_i at each sequence position i (Eq. 5.1), the proposed decoder model predicts the labels for select sequence positions $\{l_4, l_6, ..., l_i\}$.

THRE and DTTHRE can be used as elaborated in Algorithm 1. First, the diagnostic codes and primary diagnoses for each visit are extracted from the temporal health records H_p. Second, an auto-encoder AE is trained to generate the visit-level embeddings. Third, the THRE sequences (Eq. 5.2) consisting of AGE, SEX, DELTA_TIME, and VISIT medical events are constructed. Finally, DTTHRE (Eq. 5.3) is trained on the THRE sequences and primary diagnoses.

5.3 Experimental Evaluation

An important step towards developing the medical imaging cancer risk assessment model was to evaluate if embedding the elapsed time between visits impacts the predictive performance of a proof-of-concept DTTHRE model. In this section, the implementation and evaluation of two diagnostic models (Med-BERT and DTTHRE) are described. DTTHRE is the proposed decoder-only transformer (Algorithm 1) that predicts patients' primary diagnosis for each visit by analyzing their medical histories encoded using THRE. In contrast, Med-BERT (Rasmy et al., 2021) is based on the commonly used encoder-only transformer and predicts patients' primary diagnosis for their final visit by analyzing their medical histories encoded using diagnostic-level encoding.

Algorithm 1 Training DTTHRE model on THRE sequences.

Require: Patient electronic health records $H_0 ... H_P$

Ensure: THRE sequence $THRE_0 ... THRE_P$, autoencoder AE, $DTTHRE$ model, and vocabulary V

▷ Extract diagnoses and primary diagnosis from H

for $p := 0$ to P **do**
  for $n := 1$ to n_p **do**
    v_{all}.append(merge(d_p^n, m_p^n))
    l[p].append(Primary diagnosis in visit n)
  end for
end for

▷ Generate visit-level embeddings

$AE \leftarrow$ Train visit-level embedding autoencoder using v_{all}

$v_{lut} \leftarrow$ unique(v_{all})

$v_{embeddings} \leftarrow AE(v_{lut})$

▷ Construct THRE sequences (Eq. 5.2)

for $p := 0$ to P **do**
  $THRE$[p].append(Patient p age at first visit)
  $THRE$[p].append(Patient p sex)
  $THRE$[p].append(Find index of merge(d_p^1, m_p^1) in v_{lut})
  for $n := 2$ to n_p **do**
    $THRE$[p].append($t_n - t_{n-1}$)
    $THRE$[p].append(Find index of merge(d_p^n, m_p^n) in v_{lut})
  end for
end for

▷ Construct sequence vocabulary dictionary

$V \leftarrow$ Construct LUT for each unique($THRE$) element

V[Find index of v_{lut} in V, 1] $\leftarrow v_{embeddings}$

▷ Train DTTHRE (Eq. 5.3)

for epoch := 0 to 100 **do**
  for $p := 0$ to P **do**
    Forward pass $THRE$[p] through f_{dec2}
    $z_{visits} \leftarrow \{z_4, z_6, ..., z_i\}$ from last decoder layer
    $l_{pred} \leftarrow f_{c4}(z_{visits})$
    Backwards pass through f_{dec2} to update weights based on classification loss($l[p], l_{pred}$)
  end for
end for

return $THRE_0 ... THRE_P$, AE, $DTTHRE$, V

5.3.1 Data Collection and Exposure Estimation

The initial cohort consisted of 340,143 patients from the MIIDD who received at least one low-dose medical imaging scan (e.g., CT and XR) from four hospitals in Canada between May 2006 and May 2016. In this study, patients also met the following inclusion criteria: 1) Patients had at least four visits with at least one diagnosis per visit and 2) cancer patients were diagnosed $\geq$ 12 months after their first CT or XR scan. The final cohort consisted of P = 66,906 patients with 60,206 non-cancer and 6,700 cancer patients. Each patient's medical history contained demographic, health, and imaging records. Demographic data included the patient's age, sex, year, and month of the medical visit. Health records consisted of diagnostic codes in the International Statistical Classification of Diseases and Related Health Problems (ICD-10-CA) format (Canadian Institute for Health Information (CIHI), 2010). Imaging records consisted of modality (CT or XR) and body part scan (e.g., head). This study was approved by the University ethics board.

Patients' ED exposure from medical imaging was estimated using the methodology described in Chapter 4 (Boursalie et al., 2020b; Boursalie et al., 2021b). First, detailed imaging exposure records (DICOM) were collected for a subset of 1,200 patients who were a stratified random sample representative of the cohort in terms of age of first scan, sex, and body part scanned. Second, ED calculators (Kramer et al., 2010; C. Lee et al., 2015) were used to estimate ED using the DICOM headers (Boursalie et al., 2020b). Third, a PMM imputation model (Van Buuren & Groothuis-Oudshoorn, 2010) was trained to impute the ED for the cohort (Boursalie et al., 2021b). Each patient's background ED exposure was also estimated (age$*$1.7 mSv/year (Canadian Nuclear Safety Commission (CNSC), 2020)).

5.3.2 Health Records Encodings

Each patient's medical history (Fig. 5.1a) was encoded using two encoding mechanisms: 1) Med-BERT (Rasmy et al., 2021) encoding (Sect. 2.3.2) and 2) THRE (Eq. 5.2). In the Med-BERT embedding I encode: 1) 22 ICD-10-CA chapter codes, 2) background ED exposure quantized in 20 bins (range 0 to 190 mSv) (Canadian Nuclear Safety Commission (CNSC), 2020), 3) medical imaging ED exposure quantized in 78 bins (range 0 to 770 mSv), and 4) cumulative ED exposure quantized in 171 bins (range 0 to 1,700 mSv). Due to our limited dataset size (66,906 patients), the ICD-10-CA codes were encoded at the diagnostic chapter level (Table E.1) rather then the five character subcategory level used in Rasmy et al., 2021 (28 million patients). For example, ICD-10-CA codes I21.09 (Myocardial Infarction, Acute, Anterior) and I25.10 (Atherosclerotic heart disease of native coronary artery without angina pectoris) were encoded as ICD-10-CA Chapter 9 (Circulatory disease) in this study. Visits with no diagnoses (only imaging scans) were not included in the diagnostic code sequence. Instead, the medical imaging ED exposure was added to the cumulative ED total for each patient. Similarly, background, medical imaging, and cumulative ED exposures from all visits in the same month were represented as one element each and appended to the sequence after the diagnostic codes from the last visit in that month. The diagnoses, background, medical imaging, and cumulative ED exposure per visit were encoded as elements in each patient's Med-BERT sequence. The diagnoses were ordered in the sequence based on their order in each visit. The sequence position embedding was used to identify the corresponding medical visit for each element.

THRE (Section 5.2) encodes patients' medical histories as a sequence of AGE,

SEX, DELTA_TIME, and VISIT medical events. First, AGE represented the patient's age at their first visit in 0 to 100 years (101 elements) while SEX was male or female (2 elements). Second, DELTA_TIME represented the month between visits and increments from 0 to 93 months (94 elements). Third, each medical VISIT was a 160-bit embedding of a 286-bit diagnostic-level vector. Each diagnostic-level vector consisted of the ICD-10-CA chapter codes (22 bits), and quantized background (20 bits), medical imaging (78 bits), and cumulative (171 bits) ED exposure. Effective dose exposures from medical visits with no diagnoses were added to the cumulative ED total. The resulting 286-bit diagnostic-level vectors were embedded using an autoencoder into a 160-bit VISIT-level events (81,244 elements).

5.3.3 Diagnostic Prediction Transformer Models

Two diagnostic prediction models are compared: 1) Med-BERT (Sect. 2.3.2) and 2) DTTHRE (Sect. 5.2). Both models were implemented in HuggingFace PyTorch (Wolf et al., 2020). I modified the HuggingFace GPT-2 model (Sect. 5.2) to implement DTTHRE. Both models had six layers, eight attention heads per layer, and the hidden and embedding dimensions were 160-bits. The AdamWeight decay optimizer (Kingma & Ba, 2015) was used with a learning rate of 1e-5 and a dropout rate of 0.1.

The models have different optimization goals during training. Med-BERT has two training objectives (Eq. 2.1): 1) predict the primary ICD-10-CA chapter diagnosis (Table E.1) of the last visit using the patient's medical history and 2) predict the randomly masked diagnoses in the patient's medical history sequence (MLM). On the other hand, the DTTHRE model's learning objective (Eq. 5.3) is to predict the primary diagnosis for each visit based on their medical history up to that visit. To

compare DTTHRE to Med-BERT, I also evaluated DTTHRE's ability to predict the primary diagnosis of the final visit using the patient's medical history.

Med-BERT and DTTHRE were trained using stratified k-fold CV, where the dataset is split into $k = 5$ folds that contain the same proportion of class labels as the overall cohort (90% non-cancer, 10% cancer). Both models were then trained k times, using the k-1 folds (80% of the cohort) for the training set while the k^{th} fold (20% of the cohort) was used for the test set. 25% of the training set was used for the validation set. All models were trained on a GeForce RTX 2080 Ti GPU for 100 epochs with early-stopping and a batch size of 1. The models were not pretrained and fine-tuned to prevent the dataset (66,906 patient histories) from being further split. Results are summarized as mean ± standard error (SE).

5.3.4 Evaluation of Diagnostic Prediction Models

Table 5.1 shows the characteristics of the cohort and encoding mechanisms. The patients in this cohort had an average of 8 ± 0.02 medical visits with an average of 2 ± 0.003 diagnoses per visit along with their background, medical imaging, and total ED exposure. The Med-BERT diagnostic-level encoding resulted in long sequences (24.20 ± 0.07 elements per patient) which increases the search space for Med-BERT. On the other hand, THRE encoded the medical histories using shorter sequences (7.99 ± 0.03 events per patient) that incorporates more properties of the health records (age, sex, and time between visits) compared to Med-BERT. THRE resulted in a smaller search space compared to diagnostic-level encoding. However, Med-BERT encodes the visit order using the BERT's position embedding while THRE does not. The vocabulary for Med-BERT (281) was also smaller than THRE (81,440) and Rasmy et

Table 5.1: Characteristics of the cohort and encoding mechanisms (Mean ± SE). ©2021 IEEE

	Med-BERT	**DTTHRE**
Transformer	BERT (Devlin, Chang, Lee, & Toutanova, 2019)	Modified Decoder (Alg. 1)
Encoding	Med-BERT (Rasmy, Xiang, Xie, Tao, & Zhi, 2021)	THRE (Sect. 5.2)
Layers	6 layers (8 attention heads per layer)	
Learning objective	1) Predict final visit primary diagnosis 2) Predict randomly masked elements (MLM)	1) Predict each visit primary diagnosis
Cohort size	66,906	
Avg. visits per patient	8.00 ± 0.020 (Max: 96, Min 4)	
Avg. Dx per visit	2.27 ± 0.003 (Max: 25, Min 1)	
Avg. elements per patient	24.20 ± 0.07 (Max: 398, Min: 6)	14.53 ± 0.03 (Max: 193, Min: 9)
Max sequence length	400	195
Vocabulary	Dx (22), Background (20)/ imaging (78)/ total ED (171)	Age (100), Sex (2), Elapsed time (94), Visits (81,244)
Vocabulary size	291	81,400
Age	N	Y
Sex	N	Y
Diagnoses	Y	Y
Background ED	Y	Y
Imaging ED	Y	Y
Total ED	Y	Y
Visit representation	Position label denotes visits	Visit-level embedding
Diagnosis order	Y	N
Time between visits	N	Y

Table 5.2: Med-BERT and DTTHRE's precision and recall performances (mean ± SE) predicting patients' primary diagnosis in their final medical visit. The average $k = 5$ cross-validation results on the test folds are shown.

ICD Chapter Code	Number of Samples (Mean ± SD)	Med-BERT Precision (%)	Med-BERT Recall (%)	DTTHRE Precision (%)	DTTHRE Recall (%)
1	318 ± 57	68.41 ± 2.94	17.02 ± 0.22	92.74 ± 1.10	73.80 ± 0.41
2	136 ± 24	48.31 ± 1.37	18.52 ± 0.65	82.18 ± 3.36	53.22 ± 0.88
3	44 ± 8	67.41 ± 1.76	36.94 ± 1.09	81.73 ± 2.76	36.13 ± 1.02
4	98 ± 17	70.76 ± 1.33	46.71 ± 0.93	61.69 ± 3.56	35.46 ± 1.36
5	123 ± 22	63.74 ± 0.97	30.03 ± 0.76	83.19 ± 2.28	60.89 ± 1.01
6	142 ± 25	57.83 ± 1.68	22.83 ± 0.44	83.67 ± 2.14	68.05 ± 0.89
7	62 ± 11	20.00 ± 8.00	0.19 ± 0.08	97.99 ± 0.24	78.56 ± 0.99
8	133 ± 24	30.36 ± 0.72	6.97 ± 0.43	98.23 ± 0.35	88.23 ± 0.34
9	595 ± 106	67.28 ± 0.66	55.95 ± 0.35	61.86 ± 1.68	69.61 ± 0.96
10	741 ± 132	40.93 ± 0.38	50.68 ± 0.42	81.43 ± 1.27	81.65 ± 0.63
11	562 ± 101	50.27 ± 1.43	33.13 ± 0.59	69.86 ± 0.85	79.56 ± 0.55
12	188 ± 33	55.91 ± 2.37	13.08 ± 0.29	89.67 ± 1.15	73.12 ± 0.29
13	545 ± 97	37.96 ± 1.22	20.53 ± 0.52	86.42 ± 1.38	82.25 ± 0.33
14	322 ± 58	49.63 ± 1.60	30.45 ± 0.56	73.15 ± 1.20	69.79 ± 0.36
15	58 ± 10	34.87 ± 0.66	31.27 ± 0.83	88.47 ± 1.10	76.07 ± 0.61
17	18 ± 3	34.00 ± 2.71	10.36 ± 1.92	85.30 ± 2.42	58.62 ± 1.79
18	$1{,}483 \pm 265$	34.75 ± 0.45	50.88 ± 0.95	78.06 ± 1.87	82.16 ± 0.76
19	$1{,}415 \pm 253$	36.60 ± 0.35	55.38 ± 0.91	83.09 ± 1.18	89.22 ± 0.34
21	458 ± 82	37.55 ± 1.33	23.34 ± 0.88	91.18 ± 1.02	73.36 ± 0.44

al., 2021 (82,603) since the ICD-10-CA codes were encoded at the diagnostic chapter level rather then the subcategory level used in Rasmy et al., 2021.

Table 5.2 and Figures E.1 and E.2 show the test performances of the Med-BERT and DTTHRE models to predict patients' final primary diagnosis. The Med-BERT implementation had an average test accuracy of $40.51 \pm 0.13\%$ predicting the primary diagnosis for each patient's final medical visit. As shown in Table 5.2 and Fig. E.1, Med-BERT learned to predict that the primary diagnosis will be from either of the

Table 5.3: DTTHRE's precision and recall performance (mean ± SE) predicting patients' primary diagnosis in each medical visit. The average $k = 5$ cross-validation results on the test folds are shown.

ICD Chapter Code	**Number of Samples (Mean ± SD)**	**DTTHRE**	
		Precision (%)	**Recall (%)**
1	$3{,}129 \pm 12$	91.91 ± 0.86	76.78 ± 0.24
2	$1{,}067 \pm 9$	83.62 ± 3.30	59.18 ± 1.14
3	517 ± 3	85.08 ± 2.41	37.72 ± 0.82
4	$1{,}062 \pm 7$	69.98 ± 3.31	42.73 ± 1.18
5	$1{,}347 \pm 13$	83.81 ± 1.76	66.91 ± 1.10
6	$1{,}553 \pm 10$	84.81 ± 1.67	67.42 ± 0.72
7	648 ± 5	98.32 ± 0.26	76.31 ± 0.49
8	$1{,}435 \pm 3$	97.15 ± 0.31	87.10 ± 0.27
9	$5{,}912 \pm 19$	63.72 ± 1.68	71.93 ± 0.83
10	$7{,}487 \pm 21$	82.48 ± 0.90	84.05 ± 0.41
11	$6{,}522 \pm 24$	71.60 ± 0.88	80.55 ± 0.46
12	$2{,}149 \pm 11$	90.45 ± 0.82	72.78 ± 0.53
13	$6{,}012 \pm 29$	87.21 ± 1.07	83.58 ± 0.40
14	$3{,}732 \pm 11$	79.23 ± 0.97	74.96 ± 0.32
15	506 ± 4	87.01 ± 1.26	79.14 ± 0.12
16	40 ± 2	88.55 ± 2.10	51.70 ± 1.32
17	224 ± 2	89.15 ± 1.01	55.05 ± 0.71
18	$15{,}349 \pm 39$	78.28 ± 1.75	82.07 ± 0.68
19	$13{,}637 \pm 32$	82.86 ± 1.24	89.00 ± 0.33
21	$4{,}829 \pm 30$	90.32 ± 0.96	72.99 ± 0.45

two most common diagnostic codes (ICD-10-CA Chapters 18 and 19 (Table E.1)). On the other hand, the DTTHRE model (THRE encoding) was successful in predicting the primary diagnosis for each patient's final medical visit with a test accuracy of $78.54 \pm 0.22\%$, as shown in Table 5.2 and Fig. E.2. DTTHRE performance was also consistent across all primary diagnoses. Furthermore, DTTHRE achieved an average test accuracy of $79.53 \pm 0.25\%$ predicting patients' primary diagnosis for each medical visit, as shown in Table 5.3 and Fig. E.3. An advantage of the proposed model was

Table 5.4: Study findings

1.	DTTHRE successfully predicted patients' health outcomes by analyzing medical records, including the elapsed time between visits, with an average accuracy of 79.53 ± 0.25%
2.	DTTHRE, a decoder-only transformer, had improved performance (78.54 ± 0.22%) compared to Med-BERT, an encoder-only transformer model from the literature, (40.51 ± 0.13%) for health outcome prediction
3.	THRE encoding representation is a promising approach to capture complex hierarchies and relationships for analysis using transformer models

that DTTHRE predicted a primary diagnosis for each patient's visit which increased the number of training examples for the model. In fact, DTTHRE's training set size was increased to $\sum_{p=0}^{P} n_p$ with no additional training time.

5.3.5 Discussion

Table 5.4 summarizes the three main findings in this chapter. First, I proposed and demonstrated a proof-of-concept DTTHRE model that successfully predicted patients' diagnoses for each medical visit with an accuracy of 79.53 ± 0.25% on a real-world dataset. In DTTHRE, I also proposed THRE which extends the music encoding representation used by C. Huang et al., 2019 to encode medical visits, including the elapsed time between visits. The DTTHRE model and THRE encoding can be used for other health prediction tasks that require the analysis of the elapsed time between visits such as the risk of hospital readmission, infection (Wiens, Guttag, & Horvitz, 2016), and mortality prediction.

Second, DTTHRE, a decoder-only transformer, had improved performance (78.54

$\pm$ 0.22%) predicting patients' primary diagnosis in their final visit compared to Med-BERT (40.51 $\pm$ 0.13%), an encoder-only transformer. An advantage of the decoder-only transformer was that the attention heads are uni-directional so the model predicts each element in the sequence using previous elements. On the other hand, the encoder-only transformer's attention heads are bi-directional which predicts each element using all the remaining elements. As a result, the decoder-only transformer predicts the primary diagnosis for each visit in a patient's medical history which increases the training set size without additional training time. To achieve the same result in an encoder-only model, we would need to present each visit as a separate training sequence which would increase the training set size by a factor of n_p. However, presenting segments of the same patient's medical sequence multiple times to the encoder-only model for each visit may bias the model.

Third, encoding mechanisms such as THRE are a promising method to represent different data types (e.g., demographic, diagnostic, and exposure from medical imaging) for analysis using transformers. Existing transformer models have been used to analyze one data type such as text (Vaswani et al., 2017) and diagnoses (Rasmy et al., 2021). Additional features such as sequence position (Vaswani et al., 2017) and age (Y. Li et al., 2020) have been incorporated into transformers by adding them to each element in the sequence. Instead of summing features for each element, C. Huang et al., 2019 encoded multiple types of music data (time, notes on/off, and velocity) as elements in a music sequence. However, encoding medical record observations such as age, sex, the elapsed time between visits, diagnostic codes, and exposure levels as elements would result in a large model vocabulary that increases the transformer's size and training time. This study demonstrates that encoding

mechanisms can embed different data types for analysis using transformers. THRE can be extended to analyze other types of observations stored in health records such as medication and laboratory history (Choi et al., 2018).

5.3.6 Limitations

This study exhibits some limitations. First, an important component of EHR is their multi-level structure (Mirtchouk, Srikishan, & Kleinberg, 2021) which was not captured by THRE. Instead, THRE encodes the diagnostic and exposure codes for each medical visit as a concatenated visit-level encodings (Algorithm 1 Step 7 - 9). For example, the primary, secondary, and tertiary diagnoses in each visit are concatenated into a visit-level embedding. As a result, DTTHRE does not learn the hierarchical relationships in EHR. Previous studies (Choi et al., 2018; Mirtchouk et al., 2021) have proposed encoding mechanisms for EHR that encodes hierarchical EHR relationships that had improved performance compared to concatenated visit-level embeddings.

Second, I trained the DTTHRE and Med-BERT models using ICD-10-CA chapters code (e.g., lung and brain cancer are assigned as cancer elements). In contrast, Y. Li et al., 2020 and Rasmy et al., 2021 trained their encoder-only diagnostic transformers (e.g., Med-BERT) on detailed ICD-10-CA diagnostic codes (e.g., lung and brain cancer are assigned their own elements). In this study, using ICD-10-CA chapter codes decreased the vocabulary size for the Med-BERT (22 elements) compared to Rasmy et al., 2021 (82,603 elements) and DTTHRE (81,400 elements). There is a need to compare DTTHRE to Med-BERT using detailed ICD-10-CA codes on a large dataset. For example, DTTHRE and Med-BERT could be evaluated on the Medical Information Mart for Intensive Care (MIMIC; Johnson et al., 2016) dataset.

Third, DTTHRE assigns each element in the THRE sequence a distinct position label i despite elements in the sequence belonging to the same visit. Instead, all elements for each medical visit are provided to the DTTHRE's modified attention heads (Fig. 5.2). On the other hand, Y. Li et al., 2020 and Rasmy et al., 2021 labeled diagnoses from the same visit (e.g., $\{0, 0, 1, 1, ..., n_p\}$), as shown in Fig. 5.1b. As a result, DTTHRE needs to learn that elements are grouped together while Med-BERT does not. There is a need to extend DTTHRE to include visit labels.

5.4 Summary

In this chapter, the design of a new model (DTTHRE) and encoding mechanism (THRE) for analyzing irregular health record histories was reported. DTTHRE predicts patients' diagnoses for each visit by analyzing their medical histories, including the elapsed times between visits. A proof-of-concept DTTHRE was evaluated using a real medical dataset (MIIDD) and compared to an existing diagnostic transformer in the literature (Med-BERT). The proof-of-concept DTTHRE model successfully predicted patients' diagnoses with improved performance compared to Med-BERT.

I conclude with the following implications for temporally-embedded transformers based on this study:

1. **Decoder-only transformers are a promising approach for analyzing datasets with temporal properties such as health records**: In this thesis, a proof-of-concept DTTHRE successfully predicted patients' primary diagnoses by analyzing a real medical dataset, including the elapsed time between visits.

2. **Decoder-only transformer's uni-directional attention heads enables the model to predict patients' diagnoses for each visit with no additional training time**: While encoder-only transformers predict the primary diagnoses for their last visit based on their medical history, the DTTHRE model predicts the primary diagnosis for each visit based on their previous medical visits. As a result, the training set size was increased by treating each visit as a separate training example for the model.

3. **Encoding representations such as THRE is a promising approach to capture complex hierarchies and relationships for analysis using transformers**: Existing NLP (Vaswani et al., 2017) and diagnostic transformer models (Rasmy et al., 2021) encode each word and diagnosis as sequence elements which increases the transformer's size and training time. Encoding representations such as THRE provide researchers a mechanism to analyze complex concepts while decreasing the transformer's size.

Chapter 6

Conclusions and Future Work

In this chapter, the thesis contributions are summarized and future work is discussed.

6.1 Summary of Contributions

The original objective for this research was to develop an exposure risk assessment model based on deep learning to monitor the cancer risks of low-dose radiation from medical imaging. During this research, I encountered two important challenges estimating patients' exposure from medical imaging and analyzing their exposure patterns over time using deep learning. First, due to technical and privacy challenges, only a representative sample of medical images to estimate patients' exposure was accessed (Fig. 1.1a). As a result, imputation models must be used to estimate the missing exposure data in the MIIDD (Fig. 1.1b). Second, transformer models lack a mechanism to analyze temporal patterns in health records (Fig. 1.1c). Thus, the scope of this thesis was narrowed to investigate the challenges of using imputation and transformer models for predictive modeling in healthcare.

6.1.1 Evaluation of Imputation Models in Deep Learning

To address the first challenge of evaluating the imputation of missing data in health records (Fig. 1.1b), I demonstrated how adversarial machine learning techniques can be used to evaluate the performance of imputation models and their subsequent impact on deep learning models. Unlike adversarial learning, the perturbation between the imputed and actual values is a result of the imputation model rather than a malicious actor. The perturbation ranges required to impact model performance were determined. I then compared the calculated perturbations ranges to the differences between mean imputed and actual values. Despite the good agreement between the mean values, the results show that the differences between the estimation methods were enough to cause model misclassification. The findings in this thesis open new research opportunities to use concepts from adversarial machine learning to improve our understanding of the impact of imputation on deep learning.

I also assessed the performance of deep learning imputation models using the RMSE evaluation metric as in previous studies (Lall & Robinson, 2021; Yoon et al., 2018). However, I found that the RMSE performance did not agree with the qualitative evaluation using histograms. Addressing this discrepancy led to a comparative analysis between RMSE and evaluation metrics in the statistical literature, including qualitative, predictive accuracy, and statistical distance. The results of this study demonstrate that qualitative, predictive accuracy, and statistical distance metrics evaluated different qualities of the deep learning imputation model's performance. As a result, previous studies (Lall & Robinson, 2021; Yoon et al., 2018) that evaluated deep learning imputation using a predictive accuracy metric (RMSE) may not have captured the overall performances of their models. The comparative analysis in

this thesis can serve as a reference to future researchers when evaluating their own deep learning imputation models.

6.1.2 Model to Analyze Temporal Health Records

Existing diagnostic transformer models (Y. Li et al., 2020; Rasmy et al., 2021) lack a mechanism to analyze the elapsed time between medical visits. To address the second challenge of analyzing the temporal characteristics of medical histories using deep learning (Fig. 1.1c), I proposed DTTHRE (Decoder Transformer for Temporally-Embedded Health Records Encoding). DTTHRE predicts patients' diagnoses by analyzing their medical histories. In DTTHRE, instead of diagnostic-level encoding, I proposed an encoding representation for health records called THRE. THRE encodes patient histories as a sequence of medical events such as age, sex, and diagnostic embedding while incorporating the elapsed time between visits.

A proof-of-concept DTTHRE model was evaluated on a real-world medical dataset (Fig. 1.1d). DTTHRE successfully predicted patients' primary diagnoses for each visit. In addition, DTTHRE predicted patients' primary diagnosis in their final visit on a real-world medical dataset with improved performance compared to an existing diagnostic transformer model in the literature (Med-BERT). DTTHRE can be used for risk assessment models that require analyzing the time between exposure events such as radiation, pollution, and pharmacokinetics.

6.2 Future Work

In this section, recommendations for future work are discussed in three areas: 1) improve the robustness of deep learning models to imprecise imputation methods, 2) extend the imputation evaluation methodology to deep learning models, and 3) extensions to the DTTHRE model.

In Chapter 3 (Fig. 1.1b), I demonstrated how techniques from adversarial learning can be used to evaluate the impact of mean imputation on deep learning. Future work can focus on extending this study to evaluate the performance of statistical and deep learning models presented in Chapter 4 (MIDAS, GAIN, and PMM) in deep learning. In addition, a growing area of interest in adversarial learning is developing defences against adversarial attacks (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2018; Ren, Zheng, Qin, & Liu, 2020). For example, Tramer et al., 2018 proposed Ensemble Adversarial Training, a technique that augments training data with perturbation examples transferred from other models. Tramer et al., 2018 demonstrated that providing the model examples of perturbations provided robustness to adversarial attacks. Similarly, Liang and Samavi, 2020 demonstrated how a deep learning defence of ensemble networks and noisy layers can provide protection against adversarial examples while retaining accuracy. Interesting future work in this direction would be extending the proposed defences for adversarial attacks (Tramer, Carlini, Brendel, & Madry, 2020) to improve the robustness of deep learning to perturbations caused by imprecise imputation models. Third, the proof-of-concept deep learning model in Chapter 3 aggregated all data over each patient's medical history (patient-level encoding). There is a need to extend the work investigating adversarial attacks in sequential transformer models (Zhu et al., 2019; J. Li, Cao, Zhang, Chen, & Tan,

2021) to improve the model's performance analyzing datasets with imputed data. Specifically, future studies can focus on evaluating the impact of imputing features that accumulate over time (e.g., exposure).

The goal in Chapter 4 (Fig. 1.1b) was to evaluate the performance of imputation models to estimate ED exposure from medical imaging. To achieve this goal, I focused on evaluating the performance of deep learning imputation models for a single feature (ED) in the MIIDD. An advantage of deep learning was that multiple missing features can be imputed using one model. There is a need to investigate how deep learning imputations perform for multiple features compared to statistical imputation models. I also identified two limitations using the existing imputation evaluation methodology (Sect. 2.2.2) to assess deep learning imputation. First, evaluating aggregated performance across multiple features with missing data may not represent the imputation model's performance for a target feature. Second, the suitability of the evaluation metrics may depend on dataset properties such as normal vs non-normal distribution, size, and proportion of missing data. As a result, there is a need for an evaluation methodology for deep learning imputation. As the first step in this direction, I am extending the imputation evaluation methodology (MIT Critical Data, 2016) to rank deep learning models on multiple features with missing data. Future work will also focus on investigating additional evaluation metrics (similar to the study by Borji, 2019 for evaluating GAN image generation) such as mean absolute percentage error, statistical distance, and integral probability metrics (Sriperumbudur, Fukumizu, Gretton, Scholkopf, & Lanckriet, 2012).

In Chapter 5 (Fig. 1.1c), I proposed the DTTHRE model and THRE to analyze medical histories, including the elapsed time between patient visits. DTTHRE can

be extended from practical and theoretical perspectives. From a practical standpoint, DTTHRE can be used for other medical prediction tasks such as single disease prediction (e.g., cancer), risk of hospital readmission, and mortality prediction. In addition, DTTHRE can be extended to analyze other types of observations stored in health records such as medication and laboratory history (Choi et al., 2018). From a theoretical standpoint, the attention heads in DTTHRE can be visualized (Vig, 2019) to investigate the factors and relationships, including the elapsed times between events, learned by the model. For example, Y. Li et al., 2020 and Rasmy et al., 2021 used visualization techniques proposed by Vig, 2019 to investigate the diagnostic patterns learned by the models. The patterns and relationships learned by the models were then validated by clinical experts. Similarly, B. Huang, Law, and Khong, 2009 visualized the attention distribution of their decoder-only transformer to investigate what previous sequence elements the model was using to generate the next musical events. Combining DTTHRE with visualization techniques provides us a mechanism to investigate the diagnostic, exposure, visit, and temporal relations learned by the model to identify medical risk factors and relationships. Visualization of the model's attention heads will also contribute to building trust in the model. Another interesting application of deep learning is forecasting patients' disease trajectories beyond their next medical visit (Fox, Ang, Jaiswal, Pop-Busui, & Wiens, 2018). Autoregressive models such as decoder-only transformers can be used to generate new elements based on an initial sequence. For example, C. Huang et al., 2019 and Radford et al., 2019 trained decoder-only transformers to generate musical and text sequences, respectively. In the music and text transformers, previous predictions were fed back into the models to predict the next element in the sequences. Similarly, I propose

to leverage the autoregressive properties of DTTHRE to generate predictive health trajectories based on a patient's initial medical history. The proposed mechanism provides researchers and health professionals a model to predict patients' disease progression. Furthermore, the autoregressive properties of DTTHRE provide researchers and health professionals a mechanism to predict the impact of clinical decisions on patients' health. For example, the DTTHRE model could predict a patient's risk of cancer with and without the medical imaging scan being considered. As a result, DTTHRE is an important step in developing the exposure risk assessment model in the DSS. To generate disease trajectory predictions, DTTHRE needs to be extended to predict the elapsed time to the next visit.

In summary, future directions can be considered in three areas: 1) improve the robustness of deep learning to imprecise imputations methods using adversarial learning, 2) evaluation methodology for deep learning imputation models, and 3) extensions to the DTTHRE model. To improve the resilience of deep learning to imprecise imputation methods, I propose to extend the evaluation of imputation to investigate defense strategies from the adversarial learning literature. Next, I propose to extend the evaluation methodology to rank deep learning imputation models on multiple features. Then, I propose to extend DTTHRE to visualize the relationships between diagnoses, including the elapsed time between visits, learned by the model. DTTHRE can also be extended to predict disease trajectories. As a major step, I am currently extending DTTHRE to assess patients' cancer risk due to low-dose radiation exposure from medical imaging.

Bibliography

Arbel, M., Zhou, L., & Gretton, A. (2021). Generalized Energy Based Models. In *International Conference on Learning Representations (ICLR-21)*. Retrieved from https://openreview.net/forum?id=0PtUPB9z6qK

Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-based Inverse Normal Transformations are Increasingly Used, but are they Merited? *Behaviour Genetics*, *39*(5). doi:10.1007/s10519-009-9281-0

Berrington de Gonzalez, A., Mahesh, M., Kim, K., Bhargavan, M., Lewis, R., Mettler, F., & Land, C. (2009). Projected Cancer Risks From Computed Tomographic Scans Performed in the United States in 2007. *Archives of Internal Medicine*, *169*(22), 2071–2077. doi:10.1001/archinternmed.2009.440

Biau, D. J., Kere is, S., & Porcher, R. (2008). Statistics in Brief: The Importance of Sample Size in the Planning & Interpretation of Medical Research. *Clinical Orthopaedics and Related Research*, *466*(9). doi:10.1007/s11999-008-0346-9

Borji, A. (2019). Pros and Cons of GAN Evaluation Measures. *Computer Vision and Image Understanding*, *179*. doi:10.1016/j.cviu.2018.10.009

Boursalie, O., Samavi, R., & Doyle, T. (2015). M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease. In *The 5th International Conference on Current & Future Trends of Information & Communication Technologies in Healthcare (ICTH-15)*. doi:10.1016/j.procs.2015.08.357

Boursalie, O., Samavi, R., & Doyle, T. E. (2018). Machine Learning and Mobile Health Monitoring Platforms: A Case Study on Research and Implementation Challenges. *Journal of Healthcare Informatics Research*, *2*(1). doi:10.1007/s41666-018-0021-1

Boursalie, O., Samavi, R., & Doyle, T. E. (2021a). Decoder Transformer for Temporally-Embedded Health Outcome Predictions. In *IEEE The 20th International Conference on Machine Learning and Applications (ICMLA-21)*.

Boursalie, O., Samavi, R., & Doyle, T. E. (2021b). Evaluation Metrics for Deep Learning Imputation Models. In *AAAI The 5th International Workshop on Health Intelligence (W3PHIAI-21)*.

Boursalie, O., Samavi, R., Doyle, T. E., & Koff, D. A. (2020a). Deep Learning Model for Cancer Risk from Low Dose Medical Imaging Radiation. In *European Congress of Radiology*. doi:10.26044/esi2020/ESI-10315

Boursalie, O., Samavi, R., Doyle, T. E., & Koff, D. A. (2020b). Using Medical Imaging Effective Dose in Deep Learning Models: Estimation & Evaluation. *IEEE Transactions on Radiation and Plasma Medical Sciences*. doi:10.1109/TRPMS.2020.3029038

Briet, J. & Harremoes, P. (2009). Properties of Classical and Quantum Jensen-Shannon Divergence. *Physical Review A*, *79*(5). doi:10.1103/PhysRevA.79.052311

Canadian Institute for Health Information (CIHI). (2010). *Canadian Coding Standards for Version 2010 ICD-10-CA and CCI*. Retrieved from https://secure.cihi.ca/free_products/cdn-coding-standards-v2018-addendum-en.pdf

Canadian Nuclear Safety Commission (CNSC). (2020). Radiation Doses. Retrieved from https://www.cnsc-ccsn.gc.ca/eng/resources/research/index.cfm

Carlini, N. & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. In *38th IEEE Symposium on Security and Privacy (S&P-17)*. doi:10.1109/SP.2017.49

Choi, E., Xiao, C., Stewart, W., & Sun, J. (2018). MiME: Multilevel Medical Embedding of EHR for Predictive Healthcare. In *32nd Conference on Neural Information Processing Systems (NeurIPS-18)*. Retrieved from https://papers.nips.cc/paper/2018/file/934b535800b1cba8f96a5d72f72f1611-Paper.pdf

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erbaum Press. doi:10.4324/9780203771587

De Man, B., Wu, M., FitzGerald, P., Kalra, M., & Yin, Z. (2015). Dose Reconstruction for Real-time Patient-specific Dose Estimation in CT. *Medical Physics*, *42*(5). doi:10.1118/1.4921066

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*. Retrieved from https://aclanthology.org/N19-1423.pdf

Dietterich, T. G. (2002). Machine Learning for Sequential Data: A Review. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition*

(SPR) and Structural and Syntactic Pattern Recognition (SSPR). doi:10.1007/3-540-70659-3_2

Donnet, S. & Samson, A. (2013). A Review on Estimation of Stochastic Differential Equations for Pharmacokinetic/Pharmacodynamic Models. *Advanced Drug Delivery Reviews*, *65*(7). doi:10.1016/j.addr.2013.03.005

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing Data: A Systematic Review of How They are Reported and Handled. *Epidemiology*, *23*(5), 729–732. doi:10.1097/EDE.0b013e3182576cdb

European Commission. (2008). European Guidance on Estimating Population Doses from Medical X-Ray Procedures. *Radiation Protection 154*. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/72d806a2-2fb4-4e4d-a845-3b276feed8eb

European Commission. (2015). Medical Radiation Exposure of the European Population. *Radiation Protection 180*. Retrieved from https://ec.europa.eu/energy/content/rp-180-medical-radiation-exposure-european-population-part-1-part-2_en

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2018). Data Augmentation Using Synthetic Data for Time Series Classification with Deep Residual Networks. Retrieved from https://arxiv.org/pdf/1808.02455.pdf

Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R., & Wiens, J. (2018). Deep Multi-Output Forecasting: Learning to Accurately Predict Blood Glucose Trajectories. In *24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-18)*. doi:10.1145/3219819.3220102

Furukawa, K., Misumi, M., Cologne, J. B., & Cullings, H. M. (2016). A Bayesian Semiparametric Model for Radiation Dose-Response Estimation. *Risk Analysis*, *36*(6), 1211–1223. doi:10.1111/risa.12513

Gonzalez, A. B. D., Apostoaei, A. I., Veiga, L. H. S., & Land, C. (2013). RadRAT: A Radiation Risk Assessment Tool for Lifetime Cancer Risk Projection. *Journal of Radiological Protection*, *32*(3). doi:10.1088/0952-4746/32/3/205

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In *International Conference on Machine Learning (ICML-15)*. Retrieved from http://arxiv.org/abs/1412.6572

Goodfellow, I., Shlens, J., & Szegedy, C. (2016). Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representation (ICLR-16)*. Retrieved from http://arxiv.org/abs/1412.6572

Guan, J., Li, R., Yu, S., & Zhang, X. (2018). Generation of Synthetic EMR Text. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM-18)*. doi:10.1109/BIBM44415.2018

Hameed, H. & Kleinberg, S. (2020). Comparing machine learning techniques for blood glucose forecasting using free-living and patient generated data. In *5th Machine Learning for Healthcare Conference (MLMC-20)*. Retrieved from https://proceedings.mlr.press/v126/hameed20a.html

Heymans, M. & Eekhout, I. (2019). *Applied Missing Data Analysis with SPSS and R Studio*. Retrieved from https://bookdown.org/mwheymans/bookmi/

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8). doi:10.1162/neco.1997.9.8.1735

Huang, B., Law, M. W.-M., & Khong, P.-L. (2009). Whole-body PET/CT scanning: Estimation of Radiation Dose and Cancer Risk. *Radiology*, *251*(1). doi:10.1148/radiol.2511081300

Huang, C., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., ... Eck, D. (2019). Music Transformer. In *International Conference on Learning Representations (ICLR-19)*. Retrieved from https://openreview.net/pdf?id=rJe4ShAcF7

Huang, M., Zolnoori, M., Shah, N. D., & Yao, L. (2018). Temporal Sequence Alignment in EHR for Patient Representation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM-18)*. doi:10.1109/BIBM.2018.8621428

Huda, W. & Mettler, F. (2011). Volume CT Dose Index and Dose-Length Product Displayed during Computed Tomography: What Good Are They? *Radiology*, *258*(1). doi:10.1148/radiol.10100297

ICRP. (2007). The 2007 Recommendations of the International Commission on Radiological Protection (ICRP). *ICRP Publication 103*, *37*(2-4). Retrieved from https://www.icrp.org/publication.asp?id=ICRP%20Publication%20103

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., ... Giovis, C. (2005). A Review and Evaluation of Intraurban Air Pollution Exposure Models. *Journal of Exposure Science and Environmental Epidemiology*, *15*(2), 185–204. doi:10.1038/sj.jea.7500388

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). MIMIC-III, A Freely Accessible Critical Care Database. *Scientific Data*, *3*(1). doi:10.1038/sdata.2016.35

Kingma, D. P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR-15)*. Retrieved from https://arxiv.org/abs/1412.6980

Kingma, D. P. & Welling, M. (2014). Auto-encoding Variational Bayes. In *International Conference on Learning Representations (ICLR-14)*. Retrieved from https://openreview.net/forum?id=33X9fd2-9FyZd

Kramer, R., Khoury, H., & Vieira, J. (2010). CALDose_X: A Software for the Calculation of Absorbed Dose to Radiosensitive Organs and for the Assessment of Radiological Risks for Patients Submitted to X-ray Radiography. *Revista Brasileira de Fisica Medica*, *4*(2). doi:10.1088/0031-9155/53/22/011

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1). doi:10.1214/aoms/1177729694

Lall, R. & Robinson, T. (2021). The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Analysis*. doi:10.1017/pan.2020.49

Lee, C., Kim, K., Bolch, W., Moroz, B., & Les, F. (2015). NCICT: A Computational Solution to Estimate Organ Doses for Pediatric & Adult Patients Undergoing CT Scans. *Journal of Radiological Protection*, *35*(4). doi:10.1088/0952-4746/35/4/891

Lee, E., Lamart, S., Little, M., & Lee, C. (2014). Database of Normalised Computed Tomography Dose Index for Retrospective Computed Tomographic Dosimetry. *Journal of Radiological Protection*, *34*(2). doi:10.1088/0952-4746/34/2/363

Li, J., Cao, J., Zhang, Y., Chen, J., & Tan, M. (2021). Learning Defense Transformers for Counterattacking Adversarial Examples. *arXiv*. Retrieved from https://arxiv.org/abs/2103.07595

Li, L., Song, Q., & Yang, X. (2019). K-means Clustering of Overweight and Obese Population Using Quantile-transformed Metabolic Data. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *12*. doi:10.2147/DMSO.S206640

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., ... Salimi-Khorshidi, G. (2020). BEHRT: Transformer for EHR. *Nature Scientific Reports*, *10*(1). doi:10.1038/s41598-020-62922-y

Liang, Y. & Samavi, R. (2020). Towards Robust Deep Learning with Ensemble Networks and Noisy Layers. *arXiv*. Retrieved from https://arxiv.org/abs/2007.01507

Lin, E. C. (2010). Radiation Risk From Medical Imaging. *Mayo Clinic Proceedings*, *85*(12). doi:https://doi.org/10.4065/mcp.2010.0260

Little, R. J. & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons. doi:10.1002/9781119013563

Liu, M., Jiang, M., Kawai, V. K., Stein, C. M., Roden, D. M., Denny, J. C., & Xu, H. (2011). Modeling Drug Exposure Data in Electronic Medical Records: An Application to Warfarin. In *American Medical Informatics Association* (Vol. 2011). Retrieved from https://europepmc.org/articles/PMC3243123

Liu, X., Liu, C., Huang, R., Zhu, H., Liu, Q., Mitra, S., & Wang, Y. (2020). Long Short-term Memory Recurrent Neural Network for Pharmacokinetic-pharmacodynamic Modeling. *International Journal of Clinical Pharmacology and Therapeutics*, *59*(2). doi:10.5414/cp203800

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR-18)*. Retrieved from https://openreview.net/forum?id=rJzIBfZAb

Mathews, J., Forsythe, A., Brady, Z., Butler, M., Goergen, S., Byrnes, G., ... Darby, S. (2013). Cancer risk in 680,000 People Exposed to Computed Tomography Scans in Childhood or Adolescence: Data Linkage Study of 11 million Australians. *BMJ*, *346*. doi:10.1136/bmj.f2360

McCollough, C., Christner, J., & Kofler, J. (2010). How Effective Is Effective Dose as a Predictor of Radiation Risk? *American Journal of Roentgenology*, *194*(4). doi:10.2214/AJR.09.4179

Melis, M., Demontis, A., Biggio, B., Brown, G., Fumera, G., & Roli, F. (2017). Is Deep Learning Safe for Robot Vision? Adversarial Examples against the iCub Humanoid. In *International Conference on Computer Vision (ICCV-17)*. Retrieved from https://arxiv.org/abs/1708.06939

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the EHR. *Nature Science Report*, *6*(1). doi:10.1038/srep26094

Mirtchouk, M., Srikishan, B., & Kleinberg, S. (2021). Hierarchical Information Criterion for Variable Abstraction. In *Machine learning for healthcare (mlhc-21)*. Retrieved from http://www.healthailab.org/papers/21MLHC.pdf

MIT Critical Data. (2016). *Secondary Analysis of EHR*. Springer Nature. doi:10.1007/978-3-319-43742-2

National Research Council. (2006). *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. National Academies Press. Retrieved from https://www.nap.edu/catalog/11340/health-risks-from-exposure-to-low-levels-of-ionizing-radiation

Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling Incomplete Heterogeneous Data using VAES. *Pattern Recognition*, *107*. doi:10.1016/j.patcog.2020.107501

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model Checking in Multiple Imputation: An Overview and Case Study. *Emerging Themes in Epidemiology*, *14*(1). doi:10.1186/s12982-017-0062-6

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization. In *Neural Information Processing Systems (NeurIPS-16)*. Retrieved from https://papers.nips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html

Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This Time with Feeling: Learning Expressive Musical Performance. *Neural Computing And Applications*, *32*(4). doi:10.1007/s00521-018-3758-9

Osei, E. & Darko, J. (2013). A Survey of Organ Equivalent and ED from Diagnostic Radiology Procedures. *International Scholarly Research Notices Radiology*. doi:http://dx.doi.org/10.5402/2013/204346

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The Limitations of Deep Learning in Adversarial Settings. In *IEEE European Symposium on Security and Privacy (EuroS&P-16)*. doi:10.1109/EuroSP.2016.36

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85). Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2016). DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. In *Advances in Knowledge Discovery and Data Mining (AKDD-16)*. doi:10.1007/978-3-319-31750-2_3

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting Healthcare Trajectories from Medical Records: A Deep Learning Approach. *Journal of Biomedical Informatics*, *69*. doi:10.1016/j.jbi.2017.04.001

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rahman, S. A., Huang, Y., Claassen, J., Heintzman, N., & Kleinberg, S. (2015). Combining Fourier and lagged k-nearest Neighbor Imputation for Biomedical Time Series Data. *Journal of Biomedical Informatics*, *58*. doi:10.1016/j.jbi.2015.10.004

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained Contextualized Embeddings on Large-scale Structured EHR for Disease Prediction. *npj Digital Medicine*, *4*(1). doi:10.1038/s41746-021-00455-y

Reiser, M., Becker, C., Nikolaou, K., & Glazer, G. (2008). *Multislice Computed Tomography*. Springer Science & Business Media. doi:10.1007/978-3-540-33125-4

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial Attacks and Defenses in Deep Learning. *Engineering*, *6*(3). doi:10.1016/j.eng.2019.12.012

Royal, H. D. (2008). Effects of Low Level Radiation - What's New? *Seminars in Nuclear Medicine*, *38*(5). doi:https://doi.org/10.1053/j.semnuclmed.2008.05.006

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3). doi:10.1093/biomet/63.3.581

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business and Economic Statistics*, *4*(1). doi:10.2307/1391390

Rubin, D. B. (1988). An Overview of Multiple Imputation. In *American Statistical Association*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.565.6832&rep=rep1&type=pdf

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning Internal Representations by Error Propagation*. University of California San Diego. Retrieved from https://apps.dtic.mil/sti/citations/ADA164453

Schafer, P. (2016). Scalable Time Series Classification. *Data Mining and Knowledge Discovery*, *30*(5). doi:10.1007/s10618-015-0441-y

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for EHR Analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5). doi:10.1109/JBHI.2017.2767063

Sinha, R. K., Pandey, R., & Pattnaik, R. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*. doi:10.1155/2018/7068349

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Scholkopf, B., & Lanckriet, G. R. (2012). On the Empirical Estimation of Integral Probability Metrics. *Electronic Journal of Statistics*, *6*. doi:10.1214/12-EJS722

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., & Sutton, C. (2017). VEE-GAN: Reducing Mode Collapse in GANS Using Implicit Variational Learning. In *Neural Information Processing Systems (NeurIPS-17)*. Retrieved from https://proceedings.neurips.cc/paper/2017/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf

Tian, X., Yin, Z., Man, B. D., & Samei, E. (2013). Projection-based Dose Metric: Accuracy Testing and Applications for CT Design. In *Proceeding of SPIE Medical Imaging 2013: Physics of Medical Imaging* (Vol. 8668). doi:10.1117/12.2008051

Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On Adaptive Attacks to Adversarial Example Defenses. In *Conference on Neural Information Processing Systems (NeurIPS-20)*. Retrieved from https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf

Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations (ICLR-18)*. Retrieved from https://openreview.net/forum?id=rkZvSe-RZ

Van Buuren, S. & Groothuis-Oudshoorn, K. (2010). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). doi:10.18637/jss.v045.i03

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press LLC. Retrieved from https://www.routledge.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781138588318

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All You Need. In *Conference on Neural Information Processing Systems (NeurIPS-17)*. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In *57th Annual Meeting of the Association for Computational Linguistics (ACL-19)*. Retrieved from https://aclanthology.org/P19-3007.pdf

Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A., & Russell, A. (2018). Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions. *Earth and Space Science*, *5*(4). doi:10.1002/2017EA000326

Vorobeychik, Y. & Kantarcioglu, M. (2018). Adversarial Machine Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *12*(3). doi:10.2200/S00861ED1V01Y201806AIM039

Wiens, J., Guttag, J., & Horvitz, E. (2016). Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach. *Journal of Machine Learning Research*, *17*(79). Retrieved from http://jmlr.org/papers/v17/15-177.html

Wiens, J., Horvitz, E., & Guttag, J. (2012). Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. In *Advances in Neural Information Processing Systems (NeurIPS-12)* (Vol. 25). Retrieved from https://proceedings.neurips.cc/paper/2012/hash/3cec07e9ba5f5bb252d13f5f431e4bbb-Abstract.html

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Empirical Methods in Natural Language Processing (EMNLP-20)*. Retrieved from https://aclanthology.org/2020.emnlp-demos.6/

Wu, M., Yin, Z., & De Man, B. (2017). Model-based Dose Reconstruction for CT Dose Estimation. *Medical Physics*, *44*(9). doi:10.1002/mp.12409

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities & Challenges in Developing Deep Learning Models Using EHR Data: A Review. *Journal of the American Medical Informatics Association*, *25*(10). doi:10.1093/jamia/ocy068

Yeh, I.-C. & Lien, C.-h. (2009). The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert System with Applications*, *36*(2). doi:/10.1016/j.eswa.2007.12.020

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. In *35th International Conference on Machine Learning (ICML-18)* (Vol. 80). Retrieved from https://proceedings.mlr.press/v80/yoon18a.html

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., & Liu, J. (2019). FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations (ICLR-19)*. Retrieved from https://openreview.net/forum?id=BygzbyHFvB

Appendix A

MIIDD Characteristics

In this appendix (Fig. 1.1a), an overview is provided of the McMaster Imaging Information and Diagnostic Dataset (MIIDD) used for the experiments reported in this thesis. Table A.1 describes the types of medical data (descriptive, diagnostic, interventions, and imaging) available in MIIDD. Next, the breakdown of records from each data source (DAD, NACRS, and PACS) is shown in Table A.2. Then, Table A.3 describes the characteristics of the patient population (340,143 patients) represented in the dataset.

Table A.1: Types of medical data available in MIIDD

Data Type	Data Source	Description
Descriptive	DAD, NACRS, PACS	ID, Age, Sex, Postal Code, Admission and Discharge Dates, Deceased Flag
Diagnostic	DAD, NACRS	International Classification of Disease (ICD-10-CA) Codes
Interventions	DAD, NACRS	Canadian Classification of Health Interventions (CCI) Codes
Imaging	PACS	Scan Modality (CT and XR), Body Part Scanned, DICOM Scanner Values

Table A.2: MIIDD data sources breakdown

		Missing Data (%)
Number of Unique Patients	340,143	-
Number of DAD Records	282,996	0
Number of NACRS Records	1,014,737	0
Number of CT and XR DI Records	2,100,223	0
Number of CT and XR DICOM Records	39,909	98.1

Table A.3: MIIDD patient population characteristics (Mean ± SE)

		Min	**Max**
Age of First Visit (Years)	44 ± 0.05	1	100
Sex (Female)	49.00% (n = 165,393)	-	-
Patients with Record of Death	3.83% (n = 13,043)	-	-
Patients with Cancer Diagnosis	11.05% (n = 37,570)	-	-
Number of Visits Per Patient	4.34 ± 1.00E-05	1	103
Number of Diagnoses Per Visit (DAD + NACRS)	1.51 ± 1.05E-06	1	16
Number of Scans Per Patient	6.70 ± 3.23E-05	1	3,010
Number of CT Scans Per Patient	2.64 ± 8.74E-06	0	414
Number of XR Scans Per Patient	5.39 ± 2.65E-05	0	2,385

Appendix B

Electronic Health Record Samples

In this appendix (Fig. 1.1a), samples of the MIIDD health records are presented. Each MIIDD record (Table B.1) contains observations from one or more of the following: 1) DAD, 2) NACRS, or 3) PACS records. DAD records (Table B.2) include all diagnoses recorded in a medical visit. NACRS records (Table B.3) include the primary diagnosis for the medical visit and all recorded interventions. PACS records (Table B.4) include imaging information (modality and body part scanned). A sample of the detailed scanner information (DICOM headers) that are used to estimate patients' ED exposure from medical imaging is also provided in Table B.5.

Table B.1: Sample of MIIDD records curated from DAD (Table B.2), NACRS (Table B.3), and PACS (Table B.4) datasets

	Data Source	DAD Record	NACRS Record	DI Record
Patient ID	DAD/NACRS/PACS	1234567	1234567	123456
Sex	DAD/NACRS/PACS	F	F	M
Age (Years)	DAD/NACRS/PACS	66	60	79
Admit Date	DAD/NACRS/PACS	April, 2008	April, 2011	Jan, 2007
Discharge Date	DAD/NACRS/PACS	April, 2008	April, 2011	Jan, 2007
Primary Dx	DAD/NACRS	I21	I25	-
Dx Instance One	DAD/NACRS	I71	I25	-
Dx Instance Two	DAD	I21	-	-
...	...	...	...	...
Dx Instance Twenty-five	DAD	-	-	-
ICD Ch. AB Flag	DAD/NACRS	0	0	-
ICD Ch. C-D48 Flag	DAD/NACRS	0	0	-
ICD Ch. D50-D89 Flag	DAD/NACRS	0	0	-
ICD Ch. E Flag	DAD/NACRS	1	0	-
ICD Ch. F Flag	DAD/NACRS	0	0	-
ICD Ch. G Flag	DAD/NACRS	0	0	-
ICD Ch. H0-H59 Flag	DAD/NACRS	0	0	-
ICD Ch. H60-H99 Flag	DAD/NACRS	0	0	-
ICD Ch. I Flag	DAD/NACRS	1	1	-
...		...	...	...
ICD Ch. R Flag	DAD/NACRS	1	0	-
ICD Ch. ST Flag	DAD/NACRS	0	0	-
ICD Ch. V-Y Flag	DAD/NACRS	0	0	-
ICD Ch. Z Flag	DAD/NACRS	0	0	-
ICD Ch. U Flag	DAD/NACRS	0	0	-
CCI Primary Intervention	NACRS	-	3IP10VX	-
Scan Modality	PACS	-	-	CT
Workload Name	PACS	-	-	CHEST

Table B.2: Sample of DAD records from a patient's medical visit. Each column represents one diagnosis

	DAD Record 1/6	**...**	**DAD Record 6/6**
Deidentified Patient ID	1234567	...	1234567
Admit Date	Apr-08	...	Apr-08
Discharge Date	Apr-08	...	Apr-08
Sex	F	...	F
Age (Years)	66	...	66
Postal Code	L8S	...	L8S
Primary Dx	I21	...	I21
Diagnosis Instance	1	...	6
Coded Diagnosis Type Description	M - Most Responsible	...	3 - Secondary
ICD10 Coded Description	(I2149) Acute subendocardial MI	...	(I100) Benign hypertension
ICD10 Coded Chapter Description	(09) Dis of the circulatory system (I00-I99)	...	(09) Dis of the circulatory system (I00-I99)
ICD10 Coded Block Description	(I20-I25) Ischaemic heart diseases	...	(I10-I15) Hypertensive diseases
ICD10 Coded Category Description	(I21) Acute myocardial infarction	...	(I10) Essential (primary) hypertension
ICD10 Coded Subcategory Description	(I214) Acute subendocardial MI	...	(N/A) Not Applicable

Table B.3: Sample of NACRS records from a patient's medical visit. Each column represents one intervention

	NACRS Record 1/2	**NACRS Record 2/2**
Deidentified Patient ID	1234567	1234567
Admit Date	Apr-11	Apr-11
Discharge Date	Apr-11	Apr-11
Sex	M	M
Age (Years)	60	60
Postal Code	L8S	L8S
Primary Dx	I2510	I2510
Primary Dx ICD10 Category Code	I25	I25
Primary Dx ICD10 Category Description	Chronic ischaemic heart disease	Chronic ischaemic heart disease
Primary Dx Subcategory Description	Atherosclerotic heart disease	Atherosclerotic heart disease
Primary Dx Description	Ath hrt dis native coron art	Ath hrt dis native coron art
Primary Intervention CCI	3IP10VX	3IP10VX
Intervention Instance	1	2
CCI Code	3IP10VX	3KG10VX
CCI Description	Xray heart w cor art lt hrt struc PTA retro	Xray art leg after intra arterial inject contrast
CCI Section Description	Dx Imaging Interventions on the Great Vessels (3ID - 3IS)	Dx Imaging Interventions on the Lower Body Vessels (3KC - 3KU)
CCI Rubric Code	3IP10	3KG10
CCI Rubric Description	Xray heart w cor art	Xray art leg
Emergency Department?	N	N

Table B.4: Sample of PACS records

Patient ID	**Scan Date**	**Sex**	**Age**	**Modality**	**Workload Name**
1234567	Jan, 2007	M	79	CT	CHEST
1234567	Jan, 2007	M	67	CT	CHEST
1234567	Jan, 2007	M	69	CR	CHEST 1 VIEW

Table B.5: Sample of DICOM headers from CT (Toshiba and GE) and XR (Canon) scanners used to estimate ED exposure from medical imaging

Parameters	**Unit**	**DICOM Tag**	**CT (Toshiba)**	**CT (GE)**	**XR (All)**
Patient ID	-	0010,0020	1234567	1234567	1234567
Date	-	0008,0020	Jan-07	Jan-07	Jan-07
Sex	-	0010,0004	M	M	M
Age	Years	0018,0060	79	67	69
Postal Code	-	0010,1040	L8S	L8S	L8S
Modality	-	0008,0060	CT	CT	CR
Study Description	-	0008,1030	CT CHEST	CT CHEST	CHEST 1 VIEW
Series Description	-	0008,103e	Body 3.0 Sagittal	STD ALG	CHEST
Series Number	-	0020,0011	10	3	1
Scan Options	-	0018,0022	HELICAL	HELICAL	-
Protocol Name	-	0018,1030	ROUTINE CHEST NO CONTRAST	5.70 CHEST PAN STUDY	-
Manufacturer	-	0008,0070	TOSHIBA	GE	Canon Inc.
Model Name	-	0008,1090	Aquilion	Discovery CT750 HD	CXDI
Peak kVp	V	0018,0060	120	120	110
CTDIvol	mGy	0018,9345	21.9	-	-
I	mA	0018,9345	237	50	50
τ	ms	0018,1150	500	699	45
SPF	mm	0018,9311	-	0.984375	-
TCW	mm	0018,9307	-	40	-
DSD	mm	0018,1110	-	946.746	1800

Appendix C

Existing Low-dose Risk Models

Figure C.1 shows the LNT, Linear-Quadratic, Supra-linear, Linear Threshold, and Hormesis cancer risk extrapolation models (National Research Council, 2006) discussed in Section 2.1. The risk models extrapolate the cancer risk from the high doses of radiation exposures recorded in atomic bomb survivors to the low levels of radiation emitted from medical devices. The proposed extrapolation models have not been experimentally validated (National Research Council, 2006).

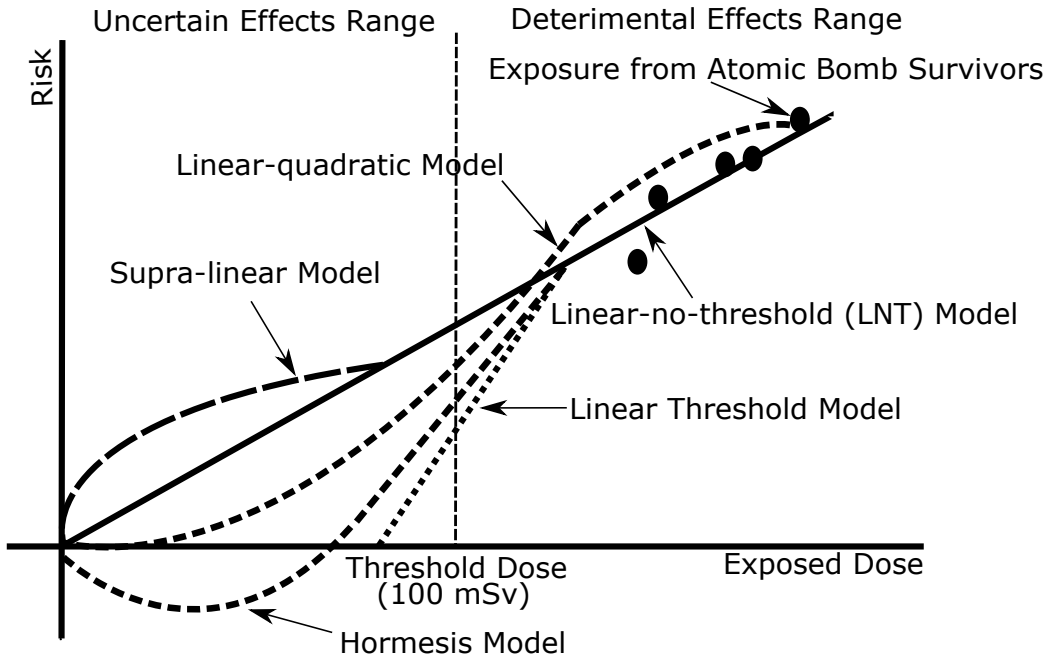


Figure C.1: The LNT, Linear-Quadratic, Supra-linear, Linear Threshold, and Hormesis cancer risk extrapolation models (Boursalie, Samavi, Doyle, & Koff, 2020a)

Appendix D

Full Imputation Histograms

In this appendix (Fig. 1.1b), the full qualitative (histogram) performances of the PMM, MIDAS, and GAIN imputation models (Chapter 4.2.2) are presented. Figures D.1 and D.2 present the imputation models' performances imputing effective dose and age, respectively. The qualitative results provide an initial check of the imputation models' performances and context to the quantitative metrics (Fig. 4.2 and Table 4.3).

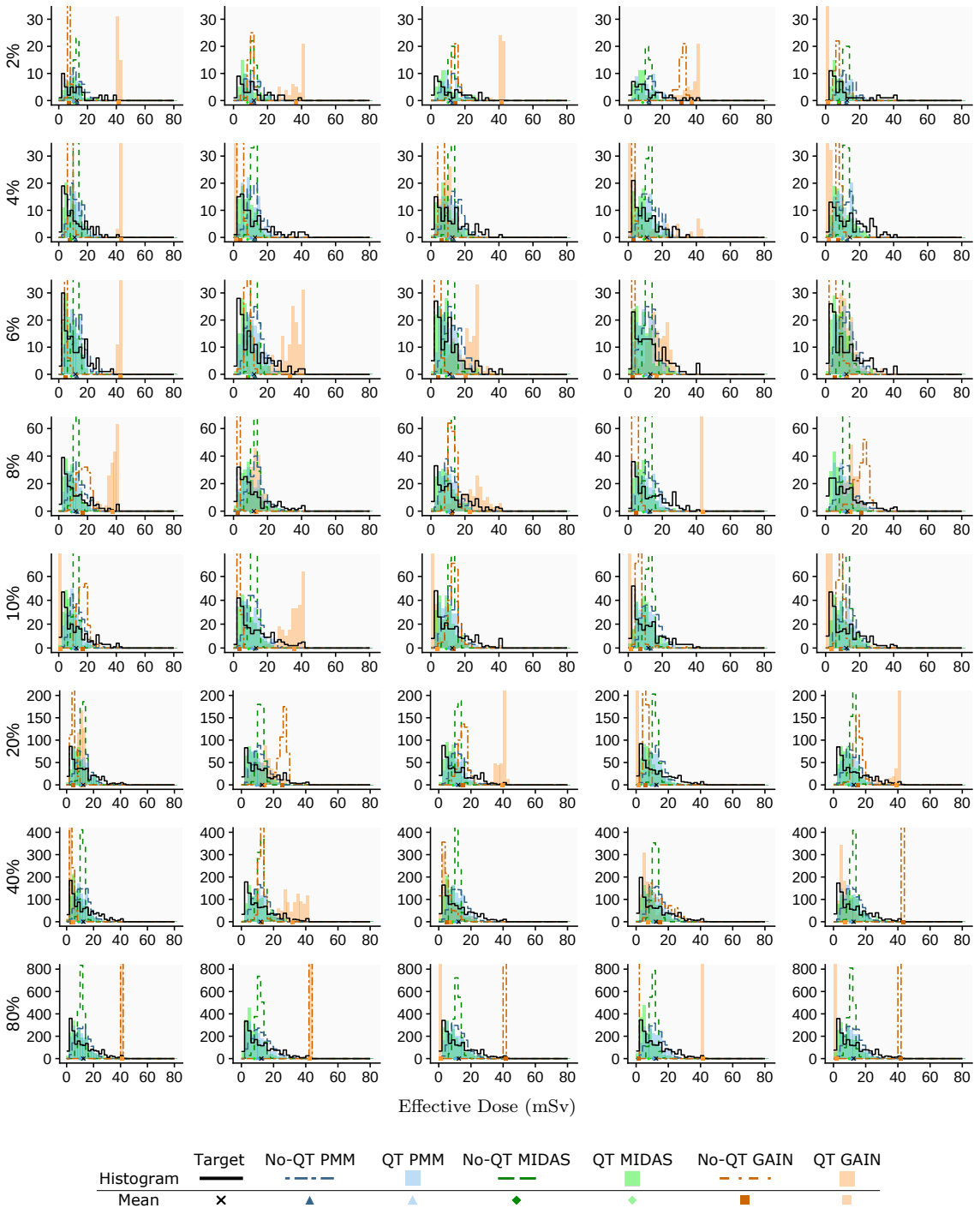


Figure D.1: Histogram of $f_{MIIDD,ED}$ imputation at 2%, 4%, 6%, 8%, 10%, 20%, 40%, and 80% missing data (rows) over five runs (columns)

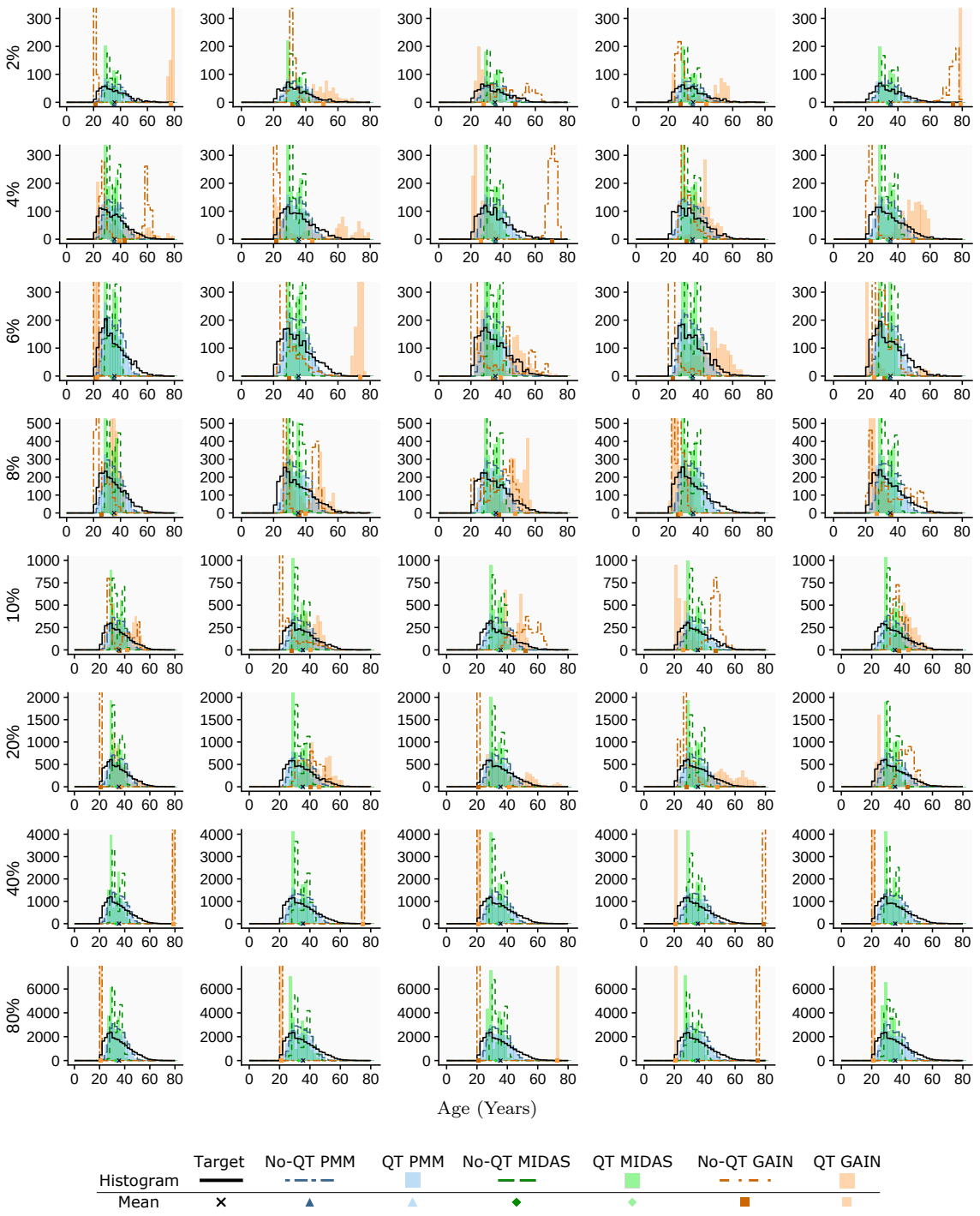


Figure D.2: Histogram of $f_{Cr,A}$ imputation at 2%, 4%, 6%, 8%, 10%, 20%, 40%, and 80% missing data (rows) over five runs (columns)

Appendix E

Confusion Matrices

In this appendix (Fig. 1.1d), the confusion matrices for the Med-BERT and DTTHRE models (Section 5.3) to predict patients' primary diagnoses are presented. Figures E.1 and E.2 show the confusion matrices of the Med-BERT and DTTHRE models respectively to predict a patient's primary diagnosis (Table E.1) in their last medical visit. DTTHRE had an improved performance (78.54 $\pm$ 0.22%) compared to Med-BERT (40.51 $\pm$ 0.13%) for health outcome prediction. Figure E.2 shows the confusion matrices of the DTTHRE model (79.53 $\pm$ 0.25%) to predict the primary diagnosis for each visit in a patient's medical history. The Med-BERT and DTTHRE models were trained with the same patient records in each $k = 5$ cross-validation fold to compare the models' performances.

Table E.1: ICD-CA-10 diagnostic chapter codes descriptions (Canadian Institute for Health Information (CIHI), 2010)

Chapter Code	**Description**
1	Certain infectious and parasitic diseases
2	Neoplasms
3	Diseases of the blood and blood-forming organs
4	Endocrine, nutritional and metabolic diseases
5	Mental, Behavioral and Neurodevelopmental disorders
6	Diseases of the nervous system
7	Diseases of the eye and adnexa
8	Diseases of the ear and mastoid process
9	Diseases of the circulatory system
10	Diseases of the respiratory system
11	Diseases of the digestive system
12	Diseases of the skin and subcutaneous tissue
13	Diseases of the musculoskeletal system and connective tissue
14	Diseases of the genitourinary system
15	Pregnancy, childbirth and the puerperium
16	Certain conditions originating in the perinatal period
17	Congenital malformations, deformations & chromosomal abnormalities
18	Codes not elsewhere classified
19	Injury, poisoning and certain other consequences of external causes
21	Factors influencing health status and contact with health services

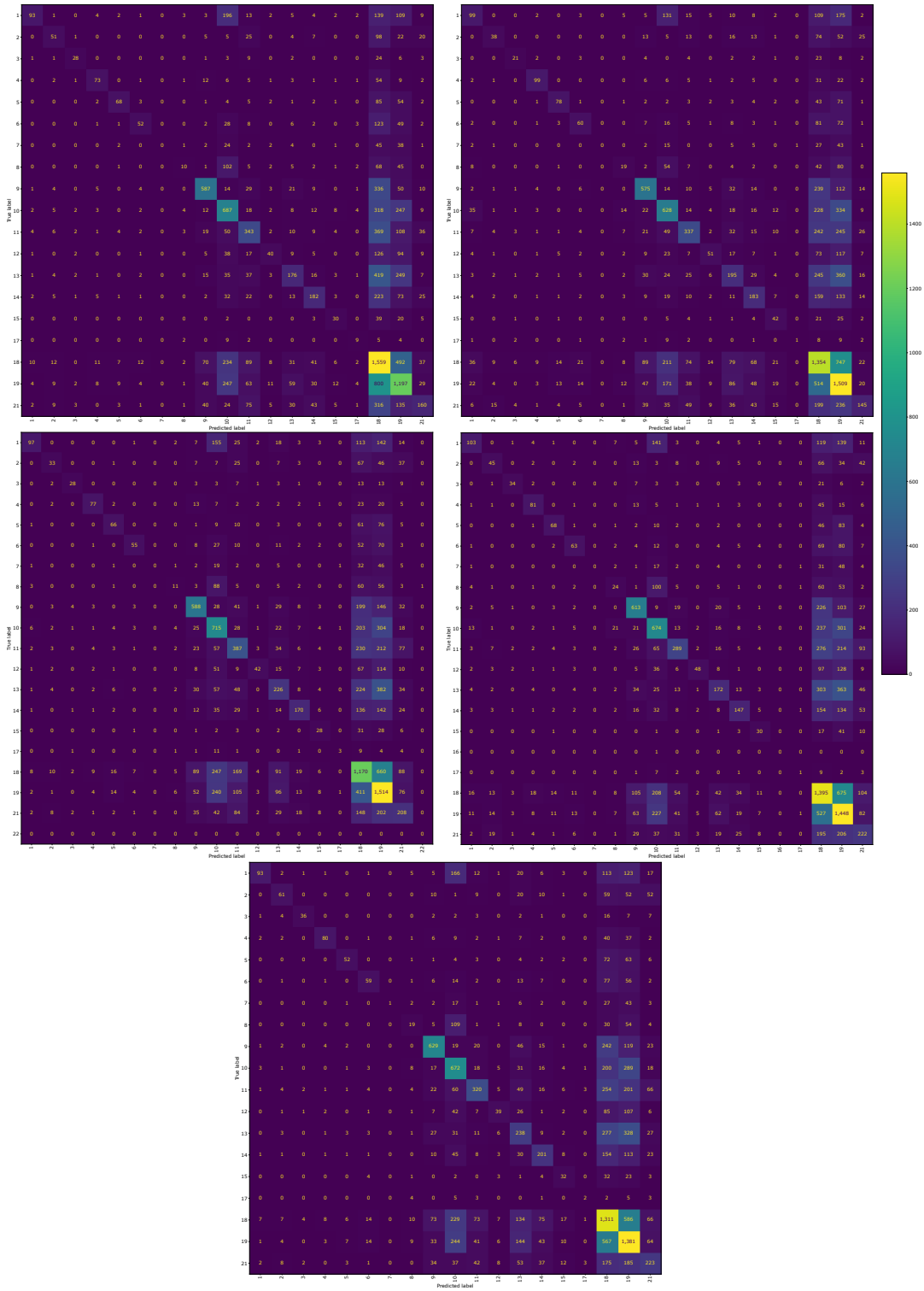


Figure E.1: Med-BERT's test confusion matrices for predicting patients' primary diagnosis in their final visit. The $k = 5$ CV results on the test folds are shown

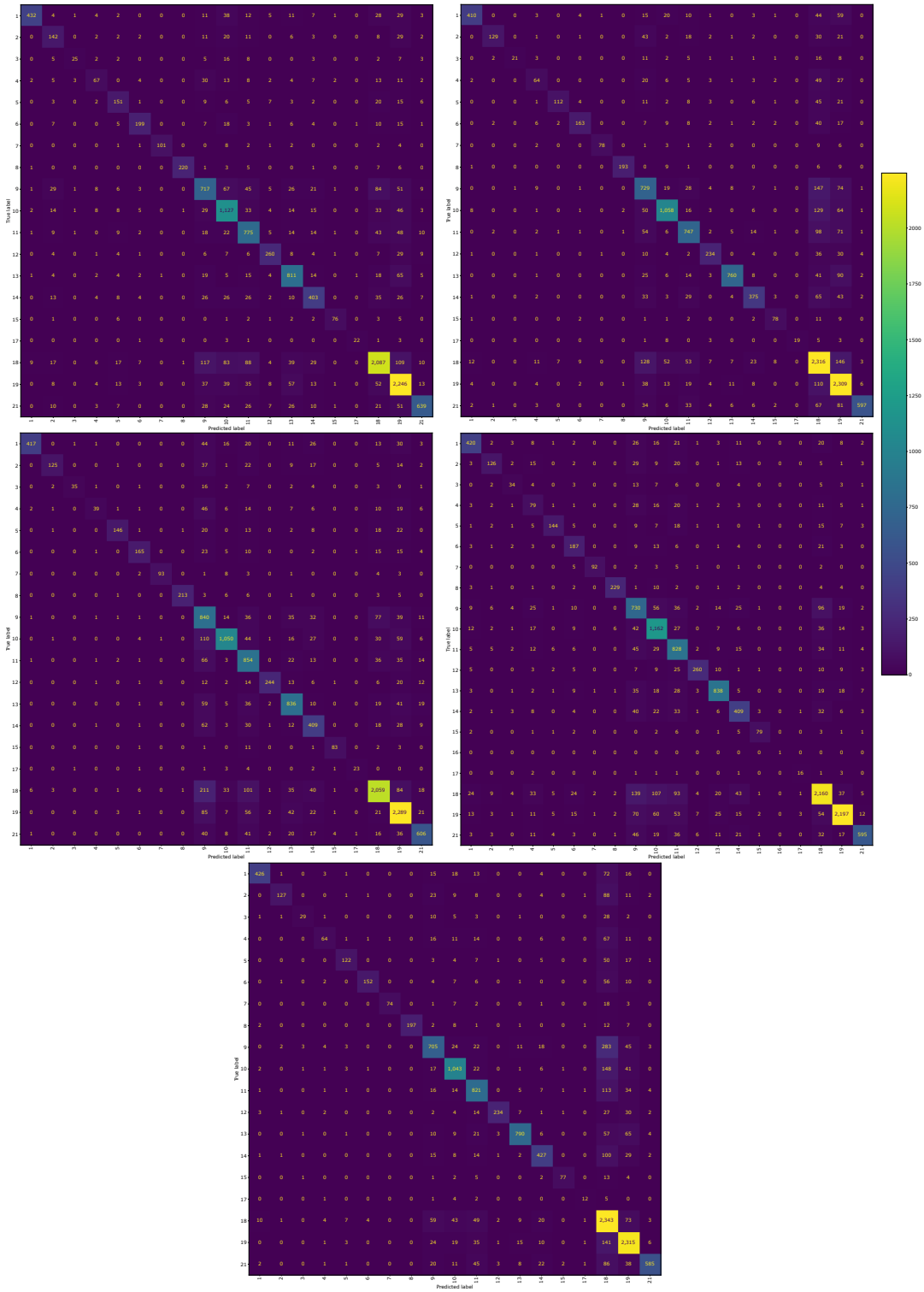


Figure E.2: DTTHRE's test confusion matrices for predicting patients' primary diagnosis in their final visit. The $k = 5$ CV results on the test folds are shown

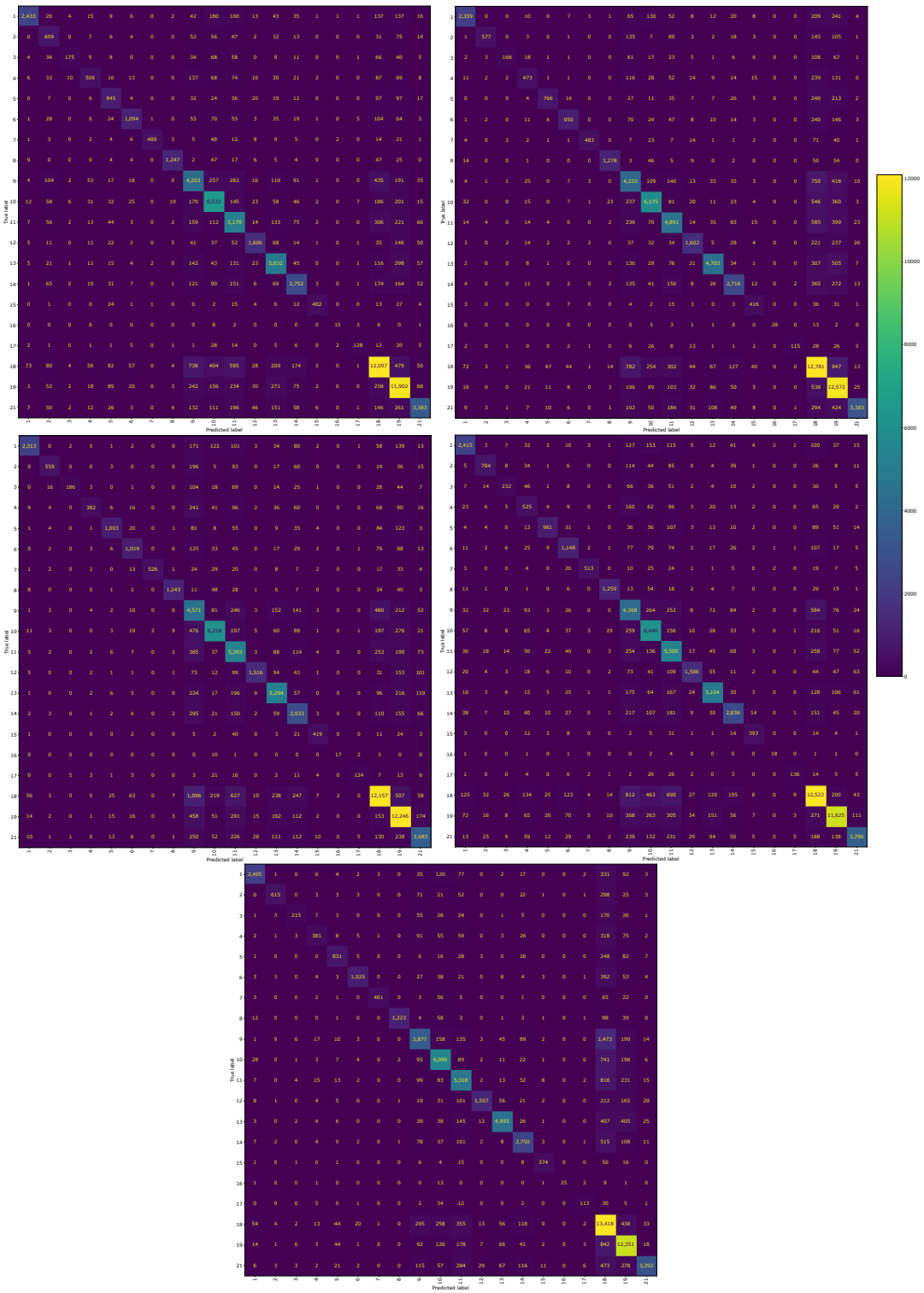


Figure E.3: DTTHRE's test confusion matrices for predicting patients' primary diagnosis in each visit. The $k = 5$ CV results on the test folds are shown