

## Representing Health Data and Medical Knowledge for Deep Learning

Hansen, Emil Riis

DOI (link to publication from Publisher):  
[10.54337/aau621651364](https://doi.org/10.54337/aau621651364)

Publication date:  
2023

Document Version  
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):  
Hansen, E. R. (2023). *Representing Health Data and Medical Knowledge for Deep Learning*. Aalborg Universitetsforlag. <https://doi.org/10.54337/aau621651364>

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

### Take down policy

If you believe that this document breaches copyright please contact us at [vbn@aub.aau.dk](mailto:vbn@aub.aau.dk) providing details, and we will remove access to the work immediately and investigate your claim.



# **REPRESENTING HEALTH DATA AND MEDICAL KNOWLEDGE FOR DEEP LEARNING**

**BY  
EMIL RIIS HANSEN**

DISSERTATION SUBMITTED 2023



**AALBORG UNIVERSITY**  
DENMARK



---

---

# Representing Health Data and Medical Knowledge for Deep Learning

---

---

Ph.D. Dissertation  
Emil Riis Hansen

Dissertation submitted August, 2023

Dissertation submitted: August, 2023

PhD supervisor:: Professor Katja Hose  
Aalborg University

PhD Co-Supervisor: Assistant Professor Tomer Sagi  
Aalborg University

PhD committee: Associate Professor Manfred Jaeger (chair)  
Aalborg University, Denmark

Associate Professor Catia Pesquita  
Falcudade de Ciencias de Universidade de Lisaboa, Portugal

Professor Yuval Shahar  
Ben-Gurion of the University Negev, Israel

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Computer Science

ISSN (online): 2446-1628  
ISBN (online): 978-87-7573-641-6

Published by:  
Aalborg University Press  
Kroghstræde 3  
DK – 9220 Aalborg Ø  
Phone: +45 99407140  
aauf@forlag.aau.dk  
forlag.aau.dk

© Copyright: Emil Riis Hansen

The author has obtained the right to include the published and accepted articles in the thesis, with a condition that they are cited, DOI pointers and/or copyright/credits are placed prominently in the references.

Printed in Denmark by Stibo Complete, 2023

# Abstract

In recent years, deep learning (DL) has grown as a powerful technology for solving complex problems in diverse domains. Thus, by leveraging the power of trainable non-linear transformations to learn intricate patterns directly from raw data in an automated way, deep learning has revolutionized how we analyze and extract new knowledge from extensive data collections.

In the meantime, the healthcare sector is entering a new digital era characterized by the growing accumulation and storage of patient biomedical information. Thus, deep learning technologies for the healthcare domain hold the prospects to complement the work of medical professionals by providing timely and accurate decision support for improved diagnostic accuracy in decision-making, reducing the time to diagnosis, and providing more personalized patient care. However, applying deep learning technologies to real-world domains is often complex. It is incredibly complex for the medical domain, where patient medical information is described by heterogeneous electronic health record (EHR) data comprising disparate modalities such as clinical images, textual descriptions, patient prescription history, laboratory tests, and many more.

Hence, in this thesis, I investigate how deep learning technologies can be used to automate how we learn from patient EHR data. First, I investigate how hierarchical medical domain taxonomies can be used to model loss functions based on patient prescription medication in a diagnosis prediction setting. Not only do I find that deep learning can be a helpful tool in predicting patients' comorbidity history based mainly on the EHR modality of patient prescription history, but also that a hierarchical loss function based on the structure of the hierarchical medical ICD taxonomy can positively benefit the task. Furthermore, I investigate the transferability and generalizability of the approach as mentioned above by applying the learned model to a large Danish EHR dataset. While the model does indeed show traits of generalizability, the EHR data from the two datasets vary in a way that challenges the direct transferring of learned models using our approach.

Subsequently, I explore the summarization of medical domain taxonomies and their application for pre-initializing node embeddings in patient graphs

connecting patients with their clinical observations. I employed graph convolution neural networks to learn to predict the patient's diagnosis codes. Results indicate that medical domain taxonomies contain rich information that can be statically extracted and used in graph-based deep learning models on EHR data. Furthermore, I investigate a novel way of structuring patient hospitalization data as sequences of medical events and utilize transformer networks for predicting the length of patient hospitalizations. Results from a hospitalization length of stay prediction problem indicate that the transformation of EHR data to sequences for integrating the temporal dependencies of the data can be beneficial for solving downstream tasks.

Lastly, I investigate how medical EHR data can be combined using multi-modal representation learning techniques to solve medical analytics problems by reviewing more than 1000 papers in the field. I find a rising trend in deep learning techniques applied to combinations of medical modalities. Furthermore, I structure medical modalities and multi-modal representation learning techniques into hierarchies based on their characteristics. I also built an explorable online analysis for researchers to dive deeper into modality combinations and their usage in medical applications based on EHR data.

Overall, while the usage of DL for EHR data is challenging, each paper in this thesis extends our knowledge by investigating novel ideas that enable us to utilize the inherent properties of EHR data in new ways. Specifically, no single data representation is always the best when working with EHR data, and the transformation of tabular data into other formats can sometimes benefit downstream tasks. Furthermore, extending DL technologies by an auxiliary component of medical domain knowledge can often be beneficial, such as hierarchical medical taxonomies.



# Resumé

Gennem de seneste år har deep learning udviklet sig til en kraftfuld teknologi til løsningen af komplekse problemer. Ved at udnytte potentialet af trænbare transformationer til automatisk at lære komplekse mønstre fra store datamængder, har deep learning revolutioneret hvordan vi analyserer og udvinder ny viden fra data.

I mellemtiden er sundhedssektoren gået ind i en ny digital æra kendetegnet ved en voksende indsamling og lagring af patienters sundhedsdata. Af netop denne grund har deep learning et stort potentiale ved at kunne supplere klinikere i deres arbejde med rettidig og præcis beslutningsstøtte til; forbedret diagnostik, reduceret diagnosetid, og personaliseret patientpleje. Det er dog ofte komplekst at skulle anvende deep learning på virkelige data. Det er især komplekst indenfor sundhedssektoren, hvor patienters sundhedsdata er kendetegnet ved heterogene data, der omfatter forskellige modaliteter som; medicinske billeder, kliniske noter, recept historik, laboratorietests og mange flere.

Jeg undersøger derfor i denne afhandling, hvordan deep learning teknologier kan bruges til at automatisere, hvordan vi lærer fra patientdata. Først undersøger jeg, hvordan hierarkiske medicinske domæne-taksonomier kan bruges til at modellere tabsfunktioner. Jeg finder at deep learning kan være et nyttigt værktøj til at modellere forudsigelsen af patienters sygdomshistorik, hovedsageligt baseret på patientens recept-historik. Herudover finder jeg at en hierarkisk tabsfunktion baseret på strukturen af den medicinske ICD-taksonomi kan have en positiv effekt på modellens ydeevne. Desuden undersøger jeg overførbareheden og generaliserbareheden af den nævnte tilgang ved at anvende den lærte model på et stort, dansk datasæt. Selvom modellen viser træk af generaliserbarhed, udgør forskellene mellem de to datasæt en udfordring for modellens direkte overførbarehed.

Dernæst udforsker jeg indlejringen af medicinske domæne-taksonomier og deres anvendelse til at initialisere node indlejringer i patientgrafer, der forbinder patienter med deres kliniske observationer. Jeg bruger graf-konvolutionelle neurale netværk til at forudsige patientens diagnosekoder. Resultaterne indikerer, at medicinske domæne-taksonomier indeholder rig

information, der kan ekstraheres statistisk og bruges i graf-baserede deep learning systemer. Desuden undersøger jeg en ny måde at strukturere patient indlæggelsesdata som sekvenser af medicinske begivenheder og anvender transformer netværk til at forudsige længden af patientindlæggelser. Resultaterne af et eksperiment, udført på et stort, dansk datasæt, indikerer, at transformationen af tabulær patientdata til sekvenser, for at integrere dataens tidsafhængigheder, kan være gavnlig for løsningen af medicinske problemer.

Endelig undersøger jeg, hvordan forskellige typer af sundhedsdata kan kombineres ved hjælp af multimodale repræsentationslæringsteknikker til at løse medicinske analytiske problemer. Ved at gennemgå mere end 1000 videnskabelige artikler, finder jeg en stigende tendens i anvendelsen af deep learning til at kombinere medicinske modaliteter. Desuden strukturerer jeg medicinske modaliteter og multimodale repræsentationslæringsteknikker i hierarkier baseret på deres egenskaber. Til slut bygger jeg et online analyseværktøj til forskere for at dykke dybere ned i modalitetskombinationer og deres anvendelse i medicinske applikationer.

Hver af mine artikler i denne afhandling udvider vores viden indenfor feltet ved at undersøge nye ideer, der forbedrer brugen af deep learning på sundhedsdata. Specifikt er ingen enkelt datarepræsentation altid den bedste, når man arbejder med sundhedsdata, og transformationen af tabeldata til andre formater kan undertiden gavne vores evne til at lære fra dataen. Desuden kan integrationen af domæneviden, såsom hierarkiske medicinske taksonomier, ofte være gavnligt i brugen af deep learning på sundhedsdata.

# Acknowledgments

First and foremost, I would like to express my gratitude to my supervisors, Katja Hose and Tomer Sagi. I am genuinely thankful for you presenting me with the opportunity to challenge life as a researcher and for your unwavering patience and support in guiding me through the journey. Your enduring patience and support in teaching and discussing my research have helped shape the researcher and person I am today. I deeply admire your dedication and meticulous attention to detail. Your guidance made this thesis possible.

Furthermore, I want to thank my good friend Kashif Rabbani for four good years of collaboration and collegiality. Our endless discussions have helped shape my mind for faster and more efficient thinking. You have helped me understand what is important, and what is not. Also, a special thanks to Matteo Lissandrini for helping me in my infant stages as a researcher and as a mental guide in the most challenging times. Truly, I have become a more happy person through our interactions. Moreover, my appreciation towards Rune Sejer Jacobsen can be only underestimated. While a PhD can be a lonely journey, our weekly work gatherings have been a light in the dark. Being in a similar field of research, you have been one of the few people who could really understand the nature of our work. I appreciate every discussion we have had, from mathematical details to philosophical and ethical discussions on the impact of our research.

Next, I want to extend my heartfelt appreciation to my incredible girlfriend, Katrine Møller Løvstad, for her unwavering support throughout this endeavor. Despite many difficulties, uncountable hours in the office, and endless rambles on my research I have made you listen to; you have always stood by my side, providing me with resolute support, comfort, motivation, and love. I greatly admire you, as this thesis would not have come to fruition without your continuous support. Thank you from the bottom of my heart.

I also sincerely thank my colleagues from the Data, Knowledge, and Web Engineering (DKW) group at Cassiopeia, Aalborg University. Your efforts in creating a pleasant work environment throughout my time here have been instrumental. You have managed to strike a perfect balance between a productive and enjoyable atmosphere, fostering great research interactions and

other (dubious) discussions. From regular coffee meetings to burger Fridays to our newest invention, PhD retreats to beautiful locations in Denmark, you have made my experience truly delightful. I feel privileged to have collaborated with such kind-hearted, dedicated, and intelligent individuals. Furthermore, I thank the administrative staff at Aalborg University for their patience and assistance; despite my frequent and sometimes repetitive inquiries, their unwavering support has been invaluable.

I also express my gratitude to all the co-authors of the papers I have contributed to during my time at the University, not just those included in this thesis. Alongside Katja and Tomer, I would like to thank Thomas Dyhre Nielsen, Thomas Mulvad, Mads Nibe Strausholm, Flemming Skjødt, Torben Bjerregaard Larsen, and Gregory Yoke Hong Lip for their valuable input and feedback throughout our collaborations. Each of them has played a significant role in shaping me as a researcher. I am genuinely grateful for the time they have dedicated to working with me during my relatively short research career thus far.

Lastly, I express my absolute appreciation to my entire family for their support during my PhD studies. Their understanding, support, and patience have been crucial in helping me navigate the long and unconventional working hours this journey has entailed.

Emil Riis Hansen  
Aalborg University, Thursday 31<sup>st</sup> August, 2023

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Resumé</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>Thesis Details</b>	<b>xiii</b>
<b>I Thesis Summary</b>	<b>1</b>
<b>Representing Health Data and Medical Knowledge for Deep Learning</b>	<b>3</b>
<b>1 Introduction</b>	<b>3</b>
1 Background and Motivation . . . . .	3
2 Thesis Structure . . . . .	10
<b>2 Domain Knowledge and Deep Learning</b>	<b>13</b>
1 Problem Motivation and Statement . . . . .	13
2 Data Homogenization and Heterogeneity . . . . .	14
3 Hierarchical Multi-label Classification . . . . .	16
4 Evaluation and Discussion . . . . .	17
5 Conclusion . . . . .	19
<b>3 Graph-based EHR Representations</b>	<b>21</b>
1 Problem Motivation and Statement . . . . .	21
2 EHR Graph Representations . . . . .	23
3 Graph Representation Learning . . . . .	24
4 Evaluation and Discussion . . . . .	26
5 Conclusion . . . . .	27

<b>4</b>	<b>Sequence-based EHR Representations</b>	<b>29</b>
1	Problem Motivation and Statement . . . . .	29
2	Patient Sequence Representations . . . . .	31
3	Event Measurements and Event Groupings . . . . .	31
4	Evaluation and Discussion . . . . .	32
5	Conclusion . . . . .	33
<b>5</b>	<b>Multi-modal Representation Learning</b>	<b>35</b>
1	Motivation and Problem Statement . . . . .	35
2	Medical Information Modality Taxonomy . . . . .	36
3	Multi-modal Representation Learning Taxonomy . . . . .	38
4	Literature Survey of MRL for Medical Analytics . . . . .	39
5	Conclusion . . . . .	40
<b>6</b>	<b>Conclusions and Future Work</b>	<b>41</b>
	References . . . . .	44
<b>II</b>	<b>Papers</b>	<b>49</b>
<b>A</b>	<b>Towards Assigning Diagnosis Codes Using Medication History</b>	<b>51</b>
1	Introduction . . . . .	53
2	Related Work . . . . .	53
3	Methods and Data . . . . .	54
3.1	Data . . . . .	54
3.2	Task - Hierarchical Multi-label Classification (HMC) . . . . .	55
3.3	Machine Learning and Loss Functions . . . . .	55
3.4	Evaluation . . . . .	57
4	Results . . . . .	58
4.1	Choice of Loss Function . . . . .	58
4.2	Discussion . . . . .	59
5	Conclusion and Future Work . . . . .	60
A	Appendix - Omitted codes and detailed results . . . . .	60
	References . . . . .	61
<b>B</b>	<b>Assigning Diagnosis Codes using Medication History</b>	<b>65</b>
1	Introduction . . . . .	67
2	Related Work . . . . .	68
3	Data and Heterogeneity . . . . .	68
3.1	MIMIC-III . . . . .	69
3.2	DNPR . . . . .	69
3.3	Homogenization . . . . .	70
3.4	Data Heterogeneity . . . . .	71
4	Hierarchical Multi-label Classification (HMC) . . . . .	74

## Contents

4.1	Machine Learning and Loss Functions . . . . .	75
5	Experimental Setup . . . . .	76
5.1	Diagnosis assignment using medication data (proposed method) . . . . .	77
5.2	Generalizability . . . . .	77
5.3	Transferability . . . . .	78
5.4	Experimentation Settings . . . . .	79
5.5	Comparison to text-based methods . . . . .	80
5.6	Baseline . . . . .	80
6	Experimental Results . . . . .	81
6.1	Diagnosis assignment using medication data (Proposed Method) . . . . .	81
6.2	Generalizability Results . . . . .	82
6.3	Transferability Results . . . . .	83
6.4	Results for text-based methods . . . . .	84
7	Discussion . . . . .	85
7.1	Proposed method . . . . .	86
7.2	Generalizability . . . . .	88
7.3	Transferability . . . . .	88
7.4	Domain Knowledge . . . . .	89
7.5	Practical Implications . . . . .	89
8	Conclusion and Future Work . . . . .	90
A	Appendix - Omitted codes and detailed results . . . . .	91
	References . . . . .	91
<b>C</b>	<b>Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies</b>	<b>95</b>
1	Introduction . . . . .	97
2	Related Work . . . . .	98
3	Initializing Graph Embeddings . . . . .	99
3.1	Multi-label Diagnosis Prediction over EHR Data . . . . .	99
3.2	Pre-initialization Using Domain Hierarchies . . . . .	101
3.3	Graph Convolution Networks . . . . .	103
4	Data . . . . .	104
5	Experiments and Results . . . . .	106
5.1	Experimental Details . . . . .	106
5.2	Results and Analysis . . . . .	108
6	Conclusion . . . . .	110
	References . . . . .	110

<b>D Patient Event Sequences for Predicting Hospitalization Length of Stay</b>	<b>113</b>
1 Introduction . . . . .	115
2 Transformer Models for EHR . . . . .	115
3 Empirical Evaluation and Results . . . . .	117
3.1 Data and Experimental Setting . . . . .	117
3.2 Results . . . . .	118
4 Conclusion . . . . .	118
References . . . . .	119
<b>E Multi-modal Representation Learning for Medical Analytics</b>	<b>121</b>
1 Introduction . . . . .	123
2 Classification Spaces . . . . .	124
2.1 Medical Analytics Utility . . . . .	125
2.2 Medical Information Modalities . . . . .	125
2.3 Multi-modal Representation Learning . . . . .	127
3 Analyzing Multi-Modal Research Contributions . . . . .	133
3.1 Modality Pairings . . . . .	135
3.2 MRL Techniques . . . . .	137
3.3 Medical Analytics Tasks . . . . .	138
4 Related work . . . . .	139
5 Conclusion . . . . .	140
5.1 Future Work. . . . .	141
References . . . . .	141



# Thesis Details

**Thesis Title:** Representing Health Data and Medical Knowledge for Deep Learning  
**Ph.D. Student:** Emil Riis Hansen  
**Supervisor:** Professor Katja Hose, Aalborg University  
**Co-Supervisor:** Assistant Professor Tomer Sagi, Aalborg University

The main body of this thesis consists of the following five papers.

- [A] Tomer Sagi, **Emil Riis Hansen**, Gregory Y. H. Lip, Torben Bjerregaard Larsen, Flemming Skjødt, Katja Hose, "Towards Assigning Diagnosis Codes using Medication History," *Proceedings of the 18th International conference on Artificial Intelligence in Medicine (AIME 2020)*, pp. 203-213, September, 2020.
- [B] **Emil Riis Hansen**, Tomer Sagi, Gregory Y. H. Lip, Torben Bjerregaard Larsen, Flemming Skjødt, Katja Hose, "Assigning Diagnosis Codes using Medication History," *In special proceedings of the 18th International conference on Artificial Intelligence in Medicine (AIME 2020)*, vol. 102308, January, 2022.
- [C] **Emil Riis Hansen**, Tomer Sagi, Katja Hose, "Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies," *In Proceedings of the 5th International Workshop on Health Data Management in the Era of AI (HeDAI@EDBT/ICDT 2023)*, vol. 3379, March, 2023.
- [D] **Emil Riis Hansen**, Thomas Dyhre Nielsen, Thomas Mulvad, Mads Nibe Strausholm, Tomer Sagi, Katja Hose, "Patient Event Sequences for Predicting Hospitalization Length of Stay," *In Proceedings of the 21st International Conference of Artificial Intelligence in Medicine (AIME 2023)*, vol. 13897, pp. 51-56, June, 2023.
- [E] **Emil Riis Hansen**, Tomer Sagi, Katja Hose, "Multi-modal Representation Learning for Medical Analytics," *Under Review*.

In addition to the above papers, I have co-authored the following two papers as part of my studies, which are not included in this thesis.

[F] **Emil Riis Hansen**, Matteo Lissandrini, Agneta Ghose, Søren Løkke, Christian Thomsen and Katja Hose, "Transparent integration and sharing of life cycle sustainability data with provenance," *In Proceedings of the 19th International Semantic Web Conference (ISWC 2020)*, 2020.

[G] Agneta Ghose, Matteo Lissandrini, **Emil Riis Hansen** and Bo Pedersen Weidema, "A core ontology for modeling life cycle sustainability assessment on the Semantic Web," *In the Journal of Industrial Ecology* (26), pp. 731-747, 2022.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the scientific papers that are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Technical Faculty of IT and Design at Aalborg University. The permission for using the published and accepted articles in the thesis has been obtained from the corresponding publishers with the conditions that they are cited and DOI pointers and/or copyright/credits are placed prominently in the references.

**Part I**

**Thesis Summary**



# Chapter 1

## Introduction

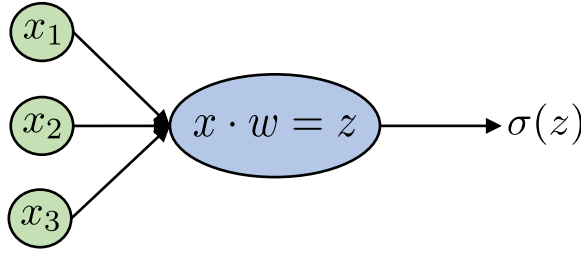
This thesis details our investigations into deep learning (DL) as a tool for automating the process of extracting insights from electronic patient health record data. The thesis is written as a collection of scientific papers. Section 1 examines the motivational aspects of the thesis and consolidates the contributions of each paper. Subsequently, Section 2 outlines the structure of the thesis.

### 1 Background and Motivation

In recent years, deep learning (DL) has emerged as a powerful tool for solving complex problems in diverse domains, revolutionizing how humans analyze and extract insights from vast volumes of data. Its unparalleled capacity to automate complex tasks has surpassed human capabilities in many aspects. At the core of DL lies the perceptron, a mathematical abstraction over the biological concept of neurons, which serves as the fundamental building block for Neural Networks (NNs). Figure 1.1 illustrates the concept of a perceptron, and Figure 1.2 illustrates the structure of a feedforward neural network consisting of multiple layers of perceptrons.

Deep learning leverages NNs, comprised of multiple layers of neurons, to effectively capture and learn intricate patterns from data through trainable non-linear transformations. By making small incremental adjustments to the weights and biases of the neurons, deep learning algorithms can automatically learn latent vector-based representations from raw data. This technology has facilitated significant advancements in difficult tasks such as natural language processing and computer vision [42, 50].

In the meantime, the healthcare sector is entering a new digital era where patient biomedical data is continuously being collected and stored from healthcare facilities, leading to ever growing collections of healthcare data. A



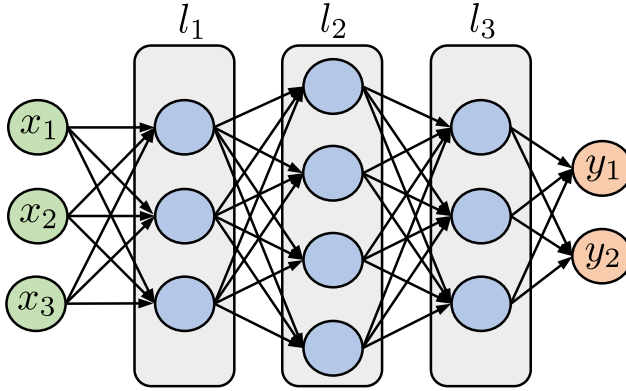
**Fig. 1.1:** Illustration of the perceptron, the building block of deep learning technologies. Given the real-valued input vector  $x$ , a linear transformation  $w$  is applied to  $x$  resulting in a single real-valued number  $z$ . Subsequently, an activation function  $\sigma$ , such as the Rectified Linear Unit or the Sigmoid function, applies non-linearity to the function [44].

patient’s healthcare data, henceforth termed electronic health records (EHR), describes individual patients’ medical history and consists of multi-modal information such as patient comorbidities, prescription medications, laboratory tests, genome data, and clinical imaging modalities as illustrated in Figure 1.3.

Given the short time allocated for medical doctors to review and analyse the vast volumes of multi-modal and high-dimensional EHR data on their patients precludes them from obtaining a unified view of their patients.

Hence, DL technologies utilizing end-to-end learning for automated pattern recognition and knowledge discovery have recently gained traction as a way to draw new insights from EHR data with the prospect of improving our ability to diagnose and treat patients [31] accurately. The following are examples of different NN architectures and their applications for solving medical analytic tasks. Recurrent Neural Networks (RNNs) are a class of NNs that excel in learning from temporal data such as text, genomes, or speech. They can consider information from prior inputs in their current computations, thus allowing them to learn from temporal dependencies. RNNs have been researched for diagnosis prediction based on clinical notes to improve the efficiency and accuracy of medical diagnosis coding [24, 46]. Convolution Neural Networks (CNNs) are a class of NNs designed to learn from data structured in grids such as images, thus making them practical for tasks such as image and video processing. They use the mathematical operation of convolutions to apply a learned filter to local features from the input while applying pooling computations to summarize local features into a low-dimensional latent representation. These operations make them robust to translational variations of objects within the input. CNNs have been used in the early-stage detection of Alzheimer’s disease [13] and for automatic classification of COVID-19 infections [32]. Mapping the human brain using

## 1. Background and Motivation

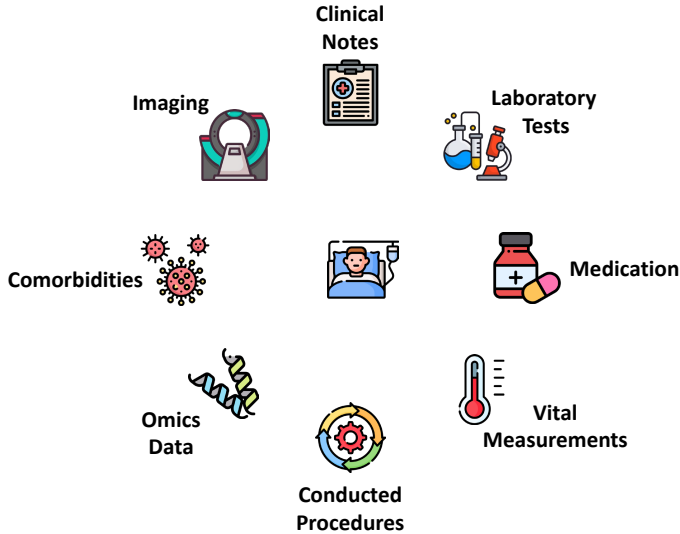


**Fig. 1.2:** Illustration of a neural network structure for learning from tabular data. Each blue dot represents a neuron. The output of each perceptron in layer 1 is used as the input for each neuron in layer 2.

graph structures of nodes representing the human brain regions and edges the morphological, functional, or structural relations between nodes. Graph Convolution Neural Network (GCNs) is a neural network that works directly with data structured as graphs consisting of nodes and relations. GCNs use permutation invariant convolution operations tailored to graph-structured data, allowing them to capture the dependencies between nodes and edges. Hence, GCNs are very useful for tasks like social network analysis, interpretation of biological systems, and recommendation. In a biomedical context, GCNs have been used in applications such as the prediction of brain disorder [4].

As DL technologies thrive on interpreting and summarizing vast volumes of data to discover novel patterns and associations that might be difficult to determine manually, DL technologies have the potential to complement and enhance the work of medical professionals by providing timely and accurate decision support that incorporates the entirety of a patient’s medical history. This can lead to benefits such as improved diagnostic accuracy in decision-making, reducing the time to diagnosis, and providing more personalized patient care [53].

However, applying DL techniques to EHR data presents complex challenges, such as missing data, data heterogeneity, and the integration of medical domain knowledge. Our work includes five papers, each building on top of the other while investigating different challenges in using DL technologies for EHR data. Paper A [41] and Paper B [22] explore how we can represent hierarchical relations directly within a loss function. Subsequently, Paper C [19] examine a methodology for embodying such hierarchical relations



**Fig. 1.3:** Illustration of patient EHR data. EHR data consist of every medical observation collected on a patient throughout the various encounters the patient has had with the healthcare sector. The data consists of multiple modalities such as textual clinical notes, medical imaging, genomics data, laboratory tests, medication administrations, and comorbidities.

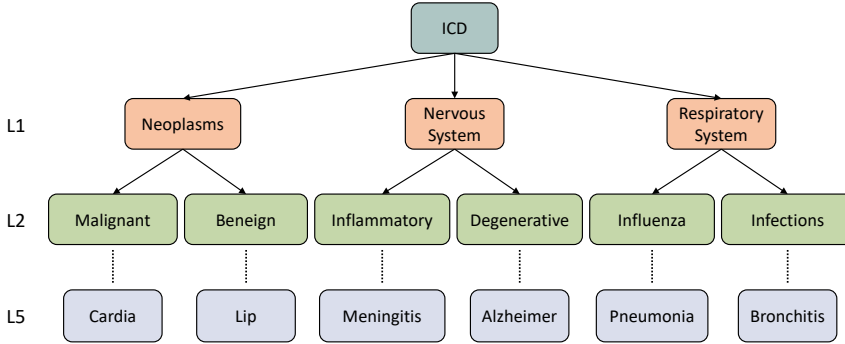
during the initialization phase of EHR graph representations. Paper D [18] explore the intricacies of modeling EHR data in a sequential format. Lastly, Paper E [20] investigate multi-modal EHR representations. The following paragraphs briefly introduce each of the papers and their relations to each other.

**Domain Knowledge and Deep Learning** Paper A [41] investigates how medical domain knowledge in the form of hierarchical medical taxonomies can be integrated into DL technologies to improve the performance of a DL-based diagnosis code prediction system. The medical domain contains an abundance of domain-specific information, including hierarchical medical taxonomies such as the International Statistical Classification of Diseases and Related Health Problems (ICD) [7] for structuring and relating diagnosis codes. Exploring new approaches to incorporate hierarchical medical taxonomies and other domain-specific information into DL models holds promise for improving their accuracy and robustness. By leveraging additional knowledge, DL models can better capture the complex interactions and dependencies within medical data, leading to more accurate predictions. Figure 1.4 illustrates the hierarchical structure of the ICD taxonomy. Paper A [41] investigates a novel hierarchical loss function for differentiating between small and large model errors using the ICD taxonomy in a diagno-



## 1. Background and Motivation

sis prediction problem. Based on the idea of multi-output hierarchical multi-label classification (HMCN) networks by Wehrmann et al. [48], we propose a novel hierarchical multi-label classification (HMC) loss function simplifying the complexity of the HMCN network structure. Specifically, we preserve the network structure of fully connected layers of neurons but extend the loss function with a hierarchical property punishing very wrong predictions more than slightly wrong predictions. Experimental results of training a model towards the task of predicting a patient’s diagnosis codes using the patient’s medication list as input to the model showed the feasibility of our approach.



**Fig. 1.4:** Illustration of a subset of the ICD hierarchical taxonomy of structuring diagnosis codes. For brevity only levels 1, 2, and 3 are shown.

Extending the work from Paper A, in Paper B [22], we investigate the model’s ability to generalize to new datasets to transfer knowledge from one dataset to another. Model generalizability and transferability are essential considerations when applying deep learning to EHR data. Models developed on a dataset gathered at one healthcare facility may not readily generalize to other populations or healthcare institutions. The performance of deep learning models must be rigorously evaluated across diverse datasets and real-world scenarios to ensure their robustness for practical usage. While our initial investigations from Paper A on a diagnosis prediction problem yielded exciting results over the publicly available MIMIC-III [26] dataset, further studies were needed to investigate the transferability and generalizability of the HMC loss function. Hence, in Paper B [22], we used a large Danish dataset of more than two million Danish patient records as a second dataset. The Danish dataset, in combination with the MIMIC-III dataset, paved the way for investigating the transferability and generalizability of our approach from Paper A. The transferability of the approach was investigated by training a deep-learning model on the MIMIC-III dataset while testing the model on the Danish dataset. The generalizability was investigated by training separate models for the datasets while comparing their performances.

The experimental results showed promising results for the generalizability of our approach. However, our approach did not yield good results in terms of transferability.

**Graph-based EHR Representations** Based on our work from Paper A and Paper B, we found that tabular data has some limitations that must be overcome to fully utilize the latent knowledge contained within EHR data. For example, simple DL models for tabular data do not handle missing data, and structural data dependencies, such as known relationships between medication and diagnosis codes, are not natively integrated. Hence, in Paper C [19], we investigate how tabular EHR data can be transformed into graph-based representations to overcome the problem of missing data while integrating the relational dependencies of the data. In the medical domain, data is not missing at random; ergo, there is a specific reason for an observation being missing [28]. While imputation techniques are commonly used to fill missing values in tabular data, we must be cautious in assuming a patient’s health status, especially if the imputed value is important for the predictive model. However, some data representations, such as graphs, naturally overcome the missing data problem while enabling the integration of multiple medical modalities. Hence, we investigate how tabular EHR data can be formatted as a graph structure based on our experience with DL-based diagnosis prediction systems in paper C [19]. Graph structures naturally capture the relational aspects of EHR data by structuring a graph as nodes representing patients and clinical observations and edges representing the relationship between a patient and a clinical concept. Using the MIMIC-III dataset, we created an EHR graph consisting of patients related to their clinical observations. We investigated how GCNs can be trained to predict patient diagnosis codes based on the events pertaining to a patient’s hospitalization. We devised a novel inductive method of pre-initializing node embeddings to increase the system’s performance by extracting the structural knowledge contained within medical domain taxonomies. Experimental results indicate that integrating domain knowledge as a pre-initialization step for graph structures could greatly benefit their performance in downstream tasks performed over the EHR graph, such as diagnosis prediction.

**Sequence-based EHR Representations** While Paper C introduced the graph-based EHR structure as a way to overcome the problem of missing data and integration of relational data, it does not naturally capture the temporal dependencies of EHR data. Hence, in Paper D [18], we investigate the transformation of EHR data into sequences of medical events pertaining to the patient.

By utilizing standardized domain vocabularies such as ICD, the Anatom-

## 1. Background and Motivation

ical Therapeutic Chemical Classification (ATC) [39], and the Nomenclature of Properties and Units (NPU) [37], we create a regulated vocabulary for describing patient sequences of medical events. Moreover, by using demographic-specific threshold values, we devise a method of incorporating measurement values as part of the event vocabulary while keeping the size of the vocabulary finite. We use a dataset of more than 47 thousand patient sequences of emergency hospitalizations from a Danish hospital to investigate sequence-based EHR representations for predicting the patients' hospitalization time, henceforth termed the patient length of stay (LOS) [45]. By creating a transformer encoder specialized in handling EHR sequences, we examined LOS prediction from the perspective of a sequence-based DL system. The results of our experiments suggest that the performance of DL systems can benefit from transforming tabular data into sequence representations to learn from the temporal aspects of the data.

**Multi-modal Representation Learning** Based on our experiences in working with EHR data in Papers A through D, we found that tabular data is easy to analyze and manage, graph representations conveniently integrate the structural dependencies of the data, and sequence representations incorporate the temporal dependencies of the data. Notwithstanding, none of these representations can easily learn from raw EHR modalities such as images, clinical notes, and omics data. Hence, in Paper E [20], we investigate the landscape of EHR modalities and their combination using multi-modal representation learning (MRL) [3] techniques to fully capture the patient's health status in a coherent way.

The world is inherently multi-modal, meaning it can only be fully understood through combining multiple senses, such as sight, sound, feeling, and taste. Likewise, patient modalities, such as clinical text, clinical imaging, and tabular observations, provide specific and complementary information on a patient's health status. Hence, the combination of medical modalities can increase the performance of DL models. While the tabular, graph, and sequence-based data representations can integrate some of these modalities, they struggle to integrate raw EHR modalities such as clinical notes, images, and timeseries data. Hence, in Paper E [20], we investigate multi-modal Representation Learning (MRL) for medical modalities, which is the area concerned with learning from multiple modalities in an automated manner. We establish a novel hierarchical taxonomy for classifying medical modalities based on their characteristics and a taxonomy for classifying MRL techniques based on their type of data combination. Subsequently, we surveyed more than 1,000 scientific papers to provide a comprehensive overview of MRL applications for medical analytics. Furthermore, we created an exploratory online analysis for researchers to investigate the survey for themselves and

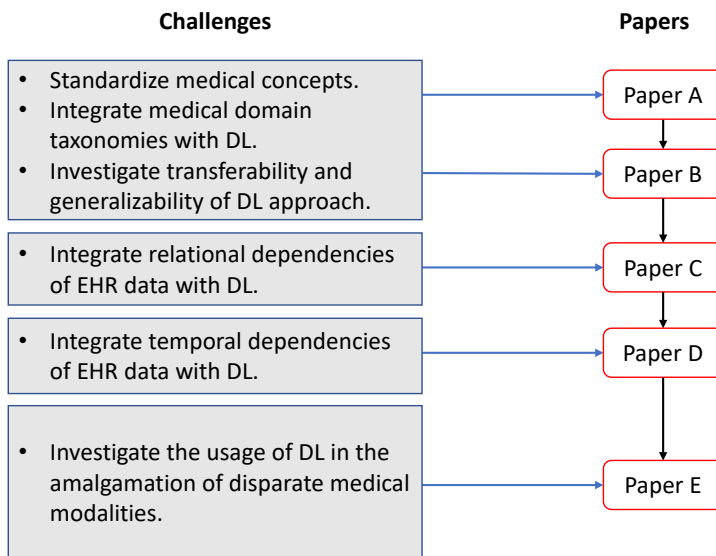
get inspiration for further development of MRL techniques and their immediate usage for specific medical applications.

## 2 Thesis Structure

The thesis is structured as follows. Part I motivates the thesis, consolidates the conducted research, and summarizes the included papers. Part II reproduces the five papers constituting this thesis, only modified by their layout to fit the thesis format. The papers can be read in any order.

Figure 1.5 illustrates the five papers in the thesis, their relations to each other, and the challenges each paper sets out to examine. Paper A and Paper B, as summarized in Chapter 2, present our efforts on the integration of the ICD hierarchical domain taxonomy in DL as a novel loss function to improve the performance of a diagnosis prediction system based on patient prescription medications. Paper C, as summarized in Chapter 3, investigates how tabular EHR data can be converted to a graph representation that integrates the relational aspects of the data, while overcoming the problem of missing data. We investigate EHR graph representations on a diagnosis prediction problem using the DL technology of graph convolution neural networks. Paper D, as summarized in Chapter 4, investigates how the temporal dependencies of EHR data can be integrated into a DL-based system by converting tabular EHR data into sequence-based representations. We investigate sequence-based EHR representations on a LOS prediction problem using the DL technology of transformer encoders. In Paper E, as summarized in Chapter 5, we propose a novel hierarchical taxonomy of medical modalities and a taxonomy of multi-modal representation learning techniques. Using these taxonomies, we survey over 1000 papers combining multiple medical modalities using multi-modal representation learning techniques for medical analytics.

## 2. Thesis Structure



**Fig. 1.5:** Overview of the five papers and their relations to each other.



## Chapter 2

# Domain Knowledge and Deep Learning

This chapter describes how we train a deep learning model towards the task of diagnosis prediction based on patients prescription medication history. Furthermore, we design a novel hierarchical loss function that, based on the structure of the ICD diagnosis codes taxonomy, takes into account the extend of model errors. The content in this chapter provides an overview of Paper A [41] and Paper B [22] and partly reuse their content.

### 1 Problem Motivation and Statement

Diagnosis code assignment is a complex problem, as shown in various studies [9, 10]. Due to the problem’s difficulty, it could be advantageous to investigate automatic diagnosis code assignment as a decision-support tool for clinicians in diagnosing patients. A diagnosis prediction tool could be used as a decision support tool for clinicians in diagnosis coding and retrospective cleaning and validating erroneous register data.

While some work exists on automatic systems for diagnosis code prediction based on laboratory test results [38] and clinical discharge notes and reports [33, 47], these systems are only applicable in the presence of these data. Moreover, text-based techniques work best when applied to English text and can not be directly transferred to new languages.

While medical record digitization has allowed physicians access to more complete and detailed medical history than ever before, its use has also rendered it overwhelming to be thoroughly reviewed in the short time that a physician can devote to each patient. The problem compounds for older patients with long medical histories and multiple comorbidities [36] and with

the unconscious patient who cannot verbally summarize their preexisting conditions. Thus, first-responders and emergency-room caregivers must rely on information from the patient’s friends, relatives, or clues in the patient’s home such as a medication list. This medication list is often pointed to as a valuable source of information that can shed light on the patient’s current medical conditions. Hence, developing a model for summarizing a patient’s medical history based on their medication list would be valuable.

Using the publicly available MIMIC-III [26] from PhysioNet [15] dataset consisting of EHR Intensive Care Unit (ICU) data for 50k American hospitalizations, including vital signs, lab results, medical notes, diagnoses ascertained, and drugs administered during the hospitalization, we aim to develop a model for automatically finding patient co-morbidities based on the administered drugs. In Paper A we develop a novel loss function utilizing the structure of the ICD taxonomy for differentiating between large and small model errors, as a way of integrating domain knowledge into DL technologies. The experiments were conducted over the MIMIC-III dataset. Subsequently, to further investigate the feasibility of our approach, in Paper B we use a large Danish dataset termed the National Danish Patient Register (NDPR), consisting of more than 2 million nationwide hospitalizations, to investigate the transferability and generalizability of our method.

The remainder of this chapter summarize the contributions of Paper A and Paper B:

- First, as described in Paper B, we homogenize the MIMIC-III and NDPR diagnosis code and prescription code concepts to promote model interoperability.
- Second, as described in Paper A, using the hierarchical ICD domain taxonomy of diagnosis codes, we design a novel hierarchical loss function for learning from the extent of model errors to promote fine-grained model learning.
- Third, as presented in Paper B, we present the results of our work, including the generalizability and transferability of our approach.

## 2 Data Homogenization and Heterogeneity

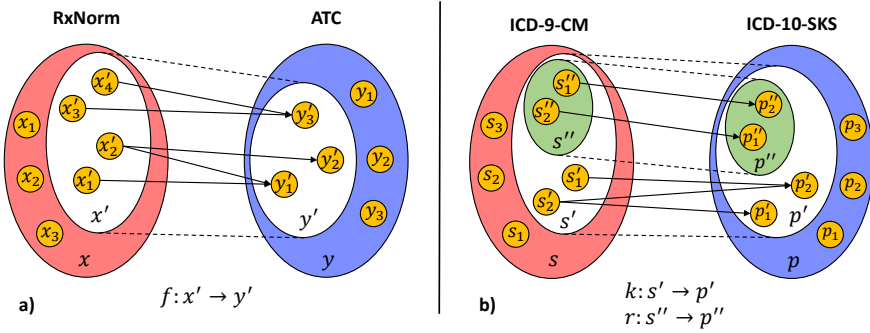
The discrepancy in medical concepts used in different datasets necessitates their homogenization before we can build a single predictive model to ingest both datasets. Hence, as we want to investigate the transferability and generalizability of our hierarchically aware loss function, we need to homogenize the MIMIC-III and NDPR datasets. MIMIC-III diagnosis code concepts are coded using a clinically modified (CM) version of ICD-9 (ICD-9-CM). ICD-9 is the 9th version of the International Classification of Diseases (ICD)



## 2. Data Homogenization and Heterogeneity

and Related Health Problems and consists of approximately 13,000 disease codes [7]. On the contrary, NDPR diagnosis code concepts are coded using the Danish Health Authority Classification System (SKS) version of ICD-10, which extends the ICD-10 taxonomy with additional branches of diseases, removes some codes not used in Denmark, and contains approximately 55,000 codes. ICD-9 has been used for disease classification of MIMIC-III throughout the years 2002-2012 patients. There is no international consensus on when to switch to newer versions of ICD, and due to the various country-specific modifications, local disease registers have been coded using different versions and modifications of ICD. Furthermore, due to the shift in granularity between earlier versions, it is impossible to fully map concepts between ICD versions earlier than ten. Forward and backward compatibility has been accounted for from the 10th version.

As the changes between ICD versions are too large, creating a bijective mapping between ICD-9-CM and ICD-10-SKS codes is impossible. However, utilizing many-to-many general equivalence mappings (GAMs)<sup>1</sup>, we managed to map 320 unique ICD-9-CM codes to 567 unique ICD-9-SKS codes in a many to many mapping. Furthermore, we created a set of 148 diagnoses that could be mapped one-to-one between ICD-9-CM and ICD-10-SKS codes. The mapping functions are illustrated in Figure 2.1b)



**Fig. 2.1:** Illustration of mapping functions for homogenization of medical concepts. **a)** illustrates the many-to-many function  $f$  for mapping between medication concepts from RxNorm and ATC. **b)** illustrates the mapping functions  $k$  and  $r$  for mapping between diagnosis concepts between ICD-9-CM and ICD-10-SKS.  $k$  is a many-to-many function mapping 567 ICD-9-CM concepts to 320 ICD-10-SKS diagnosis concepts, and  $r$  is a bijection function between ICD-9-CM and ICD-10-SKS, mapping 147 diagnosis concepts [22].

Furthermore, to disambiguate and standardize medication concepts used within the MIMIC-III dataset, we use a mapping of the MIMIC-III medication terms to standardized RxNorm [30] medication vocabulary using the the

<sup>1</sup><https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMS>

Observational Medical Outputs Partnerships (OMOP) Common Data Model (CDM) concepts [23] leaving us with 1,602 RxNorm medication codes. However, the LDPR dataset codes medication concepts using the Anatomical Therapeutic Chemical Classification (ATC) [39]. Hence, we mapped from RxNorm codes to ATC codes using the OMOP concept hierarchy, resulting in a many-to-many mapping as illustrated in Figure 2.1a). We have made all mappings available in an online appendix [21].

### 3 Hierarchical Multi-label Classification

Domain taxonomies such as ICD, ATC, and RxNorm are prevalent in the medical domain. They are used for various reasons, such as standardized coding, interoperability between healthcare systems, and efficient retrieval of medical information. Some domain taxonomies, such as ICD, are built as a hierarchical group structure based on their similar types of diseases and conditions, with the discernible diagnosis codes located as leaf nodes in the hierarchy. Exploring approaches to incorporate such domain knowledge into DL models holds promise for improving their accuracy and robustness.

Based on the global approach to multi-label classification, we investigate two loss functions for diagnosis prediction. One is suitable for multi-label classification (*ml*) where the extent of model errors are treated the same, and one integrates the hierarchical ICD taxonomy to differentiate between large and small model errors (*hml*). *ml* is the standard multi-label soft margin loss for multi-label classification as described in Equation 2.1 [52], where  $x$  is the model predictions and  $y$  is the labels.

$$\begin{aligned} \text{loss}(x, y) = & -\frac{1}{C} \sum_i y[i] \cdot \log((1 + \exp(-x[i]))^{-1}) \\ & + (1 - y[i]) \cdot \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right) \end{aligned} \quad (2.1)$$

Based on the multi-output hierarchical multi-label classification networks by Wehrmann et al. [48], we devise the *hml* loss function suitable for hierarchical multi-label classification tasks for differentiating between small and large model errors. Specifically, by preserving the network structure of fully connected neural networks, we devise an algorithm termed *roll up*, for rolling up diagnosis predictions and diagnosis labels using the ICD hierarchy. We use these in our hierarchical loss function to differentiate between small and large model errors.

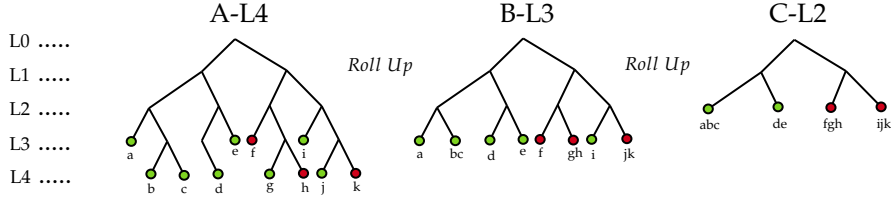
The *roll up* algorithm will roll up predictions and labels using the structure of the hierarchical ICD taxonomy. Even though a prediction might be wrong on the most specific level of aggregation, it might be right at the level above. By calculating the average over predictions of child nodes for each

#### 4. Evaluation and Discussion

parent in the next level of aggregation, we effectively get a prediction score for each diagnosis in the next level of aggregation. Subsequently, we get the labels for the next level of aggregation by calculating the max over labels of child nodes for each parent in the next level of aggregation. This procedure can be repeated until we reach the root level of the hierarchy. The roll up algorithm is illustrated in Figure 2.2.

$$\mathcal{L}_{hml} = \mathcal{L}_L + \mathcal{L}_G \quad (2.2)$$

We model  $hml$  to minimize a function comprised of two components as described in Equation 2.2. The local loss  $\mathcal{L}_L$  is calculated using Equation 2.1 and acts as the flat loss not considering the extent of errors.  $\mathcal{L}_G$  is the global loss. It is calculated using the roll up algorithm as illustrated in Figure 2.2 by rolling up the model predictions and labels one level at a time until the root level of the hierarchical ICD taxonomy. The loss from using Equation 2.1 on the new predictions for each level of hierarchical aggregation is added to the local loss.

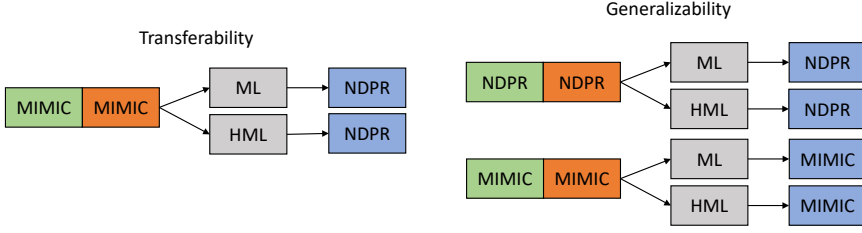


**Fig. 2.2:** Illustration of the roll up algorithm. In each iteration of the algorithm, the predictions for each node in the new level are set to the average of the model predictions over the set of child nodes. Comorbidities registered for the patient are marked with red circles, while diseases not recorded in the patient are identified with green circles [22].

## 4 Evaluation and Discussion

To investigate the feasibility, transferability and generalizability of our  $hml$  loss function, we created two experiments as illustrated in Figure 2.3. The first experiment investigates the generalizability and performance of our  $hml$  loss function by testing our approach on the MIMIC-III and LDPR datasets. Subsequently, we investigated the transferability of our approach by training a model on the MIMIC-III dataset while testing the trained model on the Danish NDPR dataset.

**Generalizability.** To homogenize the MIMIC-III and NDPR datasets, we used the inverse many-to-many functions  $k$  and inverse bijective function  $r$  to transform ICD-10-SKS codes from NDPR into ICD-9-CM codes, resulting



**Fig. 2.3:** Experiments for investigating the transferability and generalizability of our approach. Green, orange, and blue boxes represent training, validation, and test data respectively. Grey boxes represent the loss function used for training the model.

in two sets of labels for each dataset consisting of 320 and 147 diagnosis codes. Furthermore, we investigated the performance of each dataset on the top 50 and top 10 assigned diagnosis codes.

The results of the generalizability experiment indicated that incorporating domain knowledge in the form of hierarchical taxonomies for differentiating between small and large model errors into a loss function was beneficial for training a DL based system towards the task of diagnosis prediction. The experimental further indicate that, despite the MIMIC-III and NDPR datasets being heterogeneous using different ICD and medication vocabularies, both datasets have similar predictive capabilities. These findings suggest that our method is dataset agnostic. Moreover, despite the conversions between vocabulary concepts being imperfect and including many-to-many relationships, we did not find the conversion to negatively impact the performance of the models.

**Transferability.** Transferability signifies the capacity of a model to function in an environment different from the one in which it was initially trained. We assess this characteristic of our proposed method by training a model on the MIMIC-III dataset while testing it on the NDPR dataset. As for the generalizability experiments, we preprocess NDPR using the inverse mapping functions of  $k$  and  $r$  to homogenize the datasets. As in the generalizability experiment, we conducted experiments over the two sets of 320 and 147 diagnosis codes and the top 50 and top 10 assigned diagnosis codes. Furthermore, all experimental settings were investigated for the *hml* and *ml* loss functions. The model is trained and evaluated on a split of 80/20 MIMIC-III data for all transferability experiments, while the testing is conducted on the complete DNPR dataset.

The results of the transferability experiment, suggest that the heterogeneity between the datasets negatively impacts the predictive performance of our proposed method. An F1 score of 6.28% was achieved by training an

## 5. Conclusion

*hml* model on 320 diagnosis codes from the MIMIC-III dataset while testing it on the NDPR dataset. Moreover, the model could not generate any correct predictions for 229 out of 320 disease codes, yielding an F1 score of zero. This finding implies that the differences between patient data across different countries may be too significant to allow for a model with substantial transferability abilities. The differences in data collection objectives, methods, and diverse vocabulary standards for prescription and disease code hierarchies may contribute to the variance between MIMIC-III and DNPR.

However, the model transferability improves significantly when focusing on subsets of ICD-9 codes. Transferability results from the top 10 assigned diagnosis codes achieved an F1 score of 28.25. Interestingly, 4 out of the 10 ICD-9 codes produced an F1 score below 5.00, indicating that our proposed method could display high transferability for specific disease codes.

## 5 Conclusion

In Paper A and Paper B, we investigated a novel loss function that can differentiate between small and large model errors by exploiting the label relationships found in hierarchical taxonomies. Using the hierarchical ICD taxonomy of diagnosis codes, we investigated the capabilities of our loss function performance on a DL-based diagnosis prediction problem. Experiments on the American MIMIC-III dataset and the Danish NDPR datasets demonstrated our method’s superiority over using a flat loss function. However, our experiments showed limited transferability performance, indicating that a new model should be trained for each clinical setting we wish to use this technology.

In Paper A and Paper B, we investigated diagnosis prediction from the perspective of a patient’s list of prescription medications. While diagnosis prediction based solely on a patient’s prescription medication has a wide range of applications, it would be interesting to integrate more modalities into a coherent model. Hence, in the next chapter, we introduce a graph-based method of predicting a patient’s disease history to integrate more patient modalities and better capture the relational dependencies in the data.



## Chapter 3

# Graph-based EHR Representations

This chapter describes how EHR data can be modelled as graphs for subsequent patient-wise representation learning using graph convolution neural networks. Furthermore, we design and evaluate a novel node initialization method that extracts static domain knowledge from hierarchical taxonomies for inductive graph-based representation learning. The content in this chapter provides an overview of Paper C [19] and partly reuses its content.

### 1 Problem Motivation and Statement

In recent years, the world has seen a rapid rise in the collection of patient EHR data, including structured and unstructured healthcare observations. While this information presents exciting opportunities to drive progress in healthcare, the complexity and heterogeneity of the data present significant challenges for traditional machine learning techniques leveraging tabular structured data. These challenges necessitate the exploration of alternative data representations and modeling methods.

While tabular data representations are easy to manage and digest by traditional machine learning techniques, they do not fully capture the intricate relationships between patient observations. Further challenges like missing data are a big problem for tabular healthcare data. However, graph representations excel in incorporating the relational dependencies between domain concepts through their usage of nodes and multi-relational edges. As EHR graph representations can leverage the relational dependencies within data, it has recently gained traction as an input representation for the deep learning technology of graph convolution neural networks (GCNs) [17]. GCNs learn a

latent representation of graph nodes for subsequent downstream tasks, such as link prediction, whole graph classification, and node classification [49].

While much effort has been recently made in the creation of innovative GCN architectures in a model-centric way, such as the multi-relational RelationGCN [43] networks, and GraphSAGE [16] with its scalable node sampling approach, only nascent attempts have been made in data-centric integration of domain knowledge to improve the performance of graph-based DL systems. Incorporating semantically rich information into DL technologies can enhance their predictive power for solving medical tasks such as diagnosis prediction. In the context of EHR graphs, rich semantic information such as textual descriptions, hierarchical taxonomies, and uncertainty information is often inherent to the concepts described in graphs [25]; however, integrating such information has been only sparsely explored [6]. The process of pre-initializing the embeddings of graph nodes is a key method of adding domain knowledge to graphs with previous work investigating text attributes, TF/IDF scores, binary work presence vectors, node and edge degrees, and many more features for the pre-initialization of node embeddings [16, 54]. However, we are the first to investigate the latent knowledge contained within medical hierarchical taxonomies for pre-initializing node embeddings.

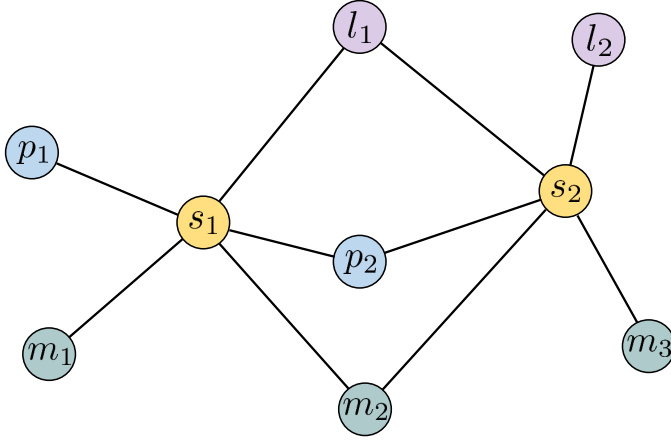
In Paper C, we explore graph representations for representing EHR data as structures consisting of nodes and edges for exploiting the intricate relationships between medical concepts. However, representing EHR data as graphs is a challenging task. Applying DL technologies such as Graph Convolution Networks (GCNs) for graph representations requires an initial latent vector representation of each node in the graph. We investigate how the structure of hierarchical medical domain taxonomies can be extracted as vector representations pertaining to the properties of domain concepts for subsequent usage in the node initialization process of EHR graphs for improved performance in the downstream task of diagnosis prediction.

By transforming EHR data from the tabular format to graph-based representations, following GCNs for learning from the EHR graph, we unveil new opportunities for learning from the steadily expanding volumes of health-care data for better decision-making support and ultimately improved patient care. The remainder of this chapter summarize the contributions of Paper C:

- First, we investigate the transformation of tabular EHR data into patient graph representations.
- Second, we devise a novel method of initializing node embeddings by extracting the latent knowledge within hierarchical medical domain taxonomies.
- Third, we present the results of our work.



## 2. EHR Graph Representations



**Fig. 3.1:** Illustration of electronic health record graph consisting of two patients  $s_1$  and  $s_2$ , two laboratory test concepts  $l_1$  and  $l_2$ , two procedure concepts  $p_1$  and  $p_2$ , and three medication concepts  $m_1$ ,  $m_2$ , and  $m_3$  [19].

## 2 EHR Graph Representations

Patient EHR data relates patients to clinical observations such as laboratory tests, vital measurements, imaging modalities, and procedures. Each observation is coded using standardized medical taxonomies such as ICD for diagnosis codes, the Nomenclature of Properties and Units (NPU) [37] codes for laboratory tests, and ATC for medication concepts. Hence, given the full set of medical concepts used to code clinical observations, we can organize a patient’s EHR data as a graph representation consisting of one node for each medical concept, one node for each patient in the dataset, and edges between nodes relating patients to their clinical observations. An example EHR graph is illustrated in Figure 3.1. The example illustrates a graph with two patients,  $s_1$  and  $s_2$ . Patient  $s_1$  is related to the set of concepts  $\{p_1, p_2, l_1, m_1, m_2\}$  and patient  $s_2$  is related to the set of concepts  $\{l_1, l_2, p_2, m_2, m_3\}$ . Furthermore,  $s_1$  and  $s_2$  are related to each other through the intersection of their respective sets of concepts, e.g.,  $\{l_1, p_2, m_2\}$ .

Graphs naturally overcome the problem of missing patient observations, as a missing observation is modeled as an edge that does not exist in the graph. Moreover, patients are naturally related to each other through the medical concepts that both patients have encountered. Given an EHR graph, the deep learning technology of graph convolution neural networks can be used to learn a latent representation of each patient that incorporates the patient’s medical history and knowledge about similar patients through their shared related medical concepts.

### 3 Graph Representation Learning

The deep learning technology of graph neural networks takes as an input a graph consisting of nodes and edges and learns from the characteristics of each node by propagating node information along the edges of the graph. GCNs thus learn how to aggregate information from the neighborhood of a node to determine a latent representation of the node that can be further used in downstream tasks such as patient diagnosis prediction. This is achieved using an *aggregation* function and an *update* function. The *aggregate* function is described in Equation 3.1. Given a node  $v$ , a user-defined function  $agg$  combines the initial feature representations of the neighborhood nodes  $\mathcal{N}(v)$  of node  $v$ .

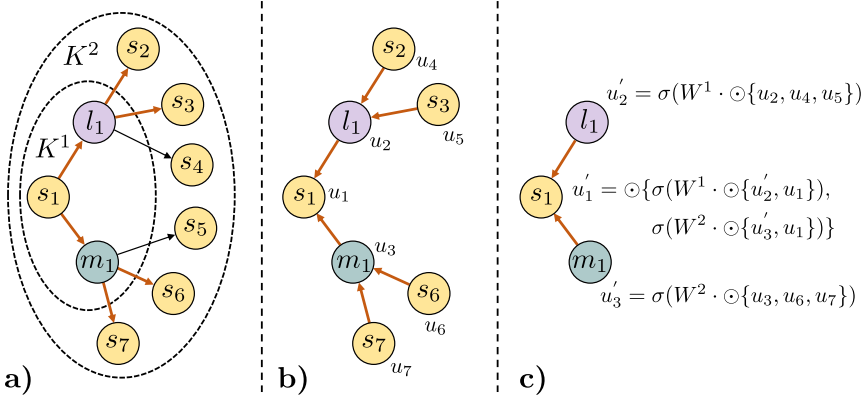
$$h_{\mathcal{N}(v)} = agg(\{h_u, \forall u \in \mathcal{N}(v)\}) \quad (3.1)$$

Subsequently, an *update* function is used to merge the initial feature representation of node  $v$  with the aggregated representation of its neighborhood as described in Equation 3.2. Given a user-defined function  $upd$ , the initial feature representation of node  $v$  is combined with the neighborhood feature representation  $\mathcal{N}(v)$ . Subsequently, a learnable non-linear transformation is applied to the output of  $upd$  through a linear transformation  $W$  and an activation function  $\sigma$ .

$$h'_v = \sigma(W \cdot upd(h_v, h_{\mathcal{N}(v)})) \quad (3.2)$$

To learn latent node embeddings in a scalable way, we use the GraphSAGE algorithm [16]. However, we modify the algorithm to work on heterogeneous graphs consisting of multiple node types to accommodate the complex nature of EHR data as illustrated in Figure 3.2. More specifically, Figure 3.2a) illustrates a 2-layer, 2-node fanout sampling strategy. The sampling strategy identifies the neighborhood  $K^1$  for patient  $s_1$ . Every sampled node in this set  $\{l_1, m_1\}$ , applies the same sampling strategy on their immediate neighborhoods  $K^2$  to further sample nodes  $\{s_2, s_3, s_6, s_7\}$ . Based on the subgraph constructed by the sampling strategy, node features from each of the 7 nodes are extracted. Finally, as illustrated in Figure 3.2c) the aggregate and update step of the graph convolution network can be used to train the network toward the downstream task of diagnosis prediction. We use a relation-specific transformation matrix  $W^i$  that operates on the average value (indicated by  $\odot$ ) of similar typed entities, as previously done in [16]. Once complete, we apply a non-linear activation function  $\sigma$  to individual convolutions. If it becomes necessary to combine features of different types, such as in the combination of  $\{u_1, u'_2, u'_3\}$ , we use the element-wise mean to integrate individual transformation.

### 3. Graph Representation Learning



**Fig. 3.2:** Illustration of the steps involved in our graph convolution process. **a)** A sampling strategy is used to define a small set of nodes. **b)** Second, node representations of the subgraph are extracted. **c)** Finally, a new latent representation for node  $s_1$  is calculated by combining its current representation with the representations of its neighborhood [19].

As the initial feature values of node embeddings serve as the starting point for all successive transformations and updates, they greatly impact the performance of GCN models. Hence, appropriately initializing these embeddings can significantly influence the ability of the model to capture and reflect the intricate relational patterns within the graph, thereby leading to more accurate predictions. Influenced by our work in using hierarchical medical taxonomies to improve the predictive capabilities of a diagnosis prediction system in Paper A and Paper B, we investigated whether hierarchical taxonomies could also be used in the pre-initialization of node feature embeddings.

We surmise that the hierarchical position of medical concepts within their respective standardized hierarchical taxonomies could contain semantically relevant information for diagnosis prediction. For example, the medication metformin (coded as A10BA02) at the top level of the ATC hierarchy indicates that it targets the alimentary tract and metabolism, the second level reveals it is used for diabetes, the third level shows it lowers blood glucose, the fourth level classifies it in the chemical subgroup of biguanides, and the final level identifies the chemical substance as metformin. Hence, if the patient has been prescribed metformin, they likely have type 2 diabetes. Integrating such information into the DL model should enhance the model's ability to learn the usage of specific medications. Similarly, hierarchies exist for surgical procedures, coded by the ICD-9 Procedures (PROC) taxonomy, which groups related procedures based on the operation site. If a patient underwent partial adrenalectomy (code 07.2), they likely had a condition related to the endocrine glands. Likewise, laboratory tests are coded using the hierarchical

LOINC concept codes, which group related laboratory tests based on their class, component, and system, which provides important information on the purpose of the laboratory test.

Using a combination of breadth-first and depth-first searches, we devised an algorithm *TreeEmb* to extract the structural knowledge of medical concepts from their hierarchical medical taxonomies. The computed concept attributes can then be leveraged to pre-initialize node embeddings of EHR concepts. Moreover, this method of creating embeddings guarantees that concepts closely linked in the tree will have more similar embeddings than those with a larger distance. Herefore, GCNs will more likely learn that clusters of closely associated concepts are used in treating the same disease, thereby reducing the epistemic uncertainty by incorporating domain knowledge.

## 4 Evaluation and Discussion

To investigate our proposed method of pre-initializing graph node embeddings using the structural knowledge contained within hierarchical medical taxonomies, we created a graph representation of the MIMIC-IV EHR dataset from PhysioNet, which contains data from 382,278 emergency care patients. The MIMIC-IV database includes laboratory results, vital measurements, determined diagnoses, administered medications, and demographic data structured as a relational database.

We convert the dataset into the OMOP CDM format using an Extract-Transform-Load (ETL) conversion process to disambiguate medical concepts. The CDM format helps standardize and clarify medical concepts, providing a bridge for future AI models to work on different datasets converted to the CDM format. In the CDM format, laboratory tests, procedures, and medications are encoded using the LOINC, ICD-9, and RxNorm taxonomies. As RxNorm is a non-hierarchical taxonomy, we link each medication concept to the hierarchical ATC medication taxonomy through its active ingredients.

We experimented with three different methods of node pre-initialization, including graphlet and edge count features [1] (**Graphlet**) [40], random-initialized node embeddings using Xavier initialization [14] (**Rand**), and our method (**FeatInit**) using the *TreeEmb* algorithm to extract the structure of hierarchical domain taxonomies. The random-initialized node embedding method serves as a transductive baseline, as the node embeddings are made trainable as part of the supervised model training phase [17]. The **Graphlet** and **FeatInit** methods are inductive in the way that original node features are not changed during the graph training phase.

We found that using *TreeEmb* embeddings to pre-initialize node embeddings led to an improvement in F1 scores compared to when node embeddings were learned during training or pre-initialized using graphlet features.

## 5. Conclusion

These findings indicate that medical domain hierarchies contain knowledge that, when integrated with DL technologies, can be beneficial for solving downstream tasks such as patient diagnosis prediction.

## 5 Conclusion

In Paper C, we investigated how tabular EHR data can be transformed into graph representations to integrate the relational dependencies in the data. Subsequently, we trained a GCN model toward the task of diagnosis prediction by learning a latent embedding for each patient. Furthermore, we devise a novel node pre-initialization method for adding domain knowledge to the initial graph representation using the latent knowledge contained within hierarchical medical domain taxonomies. Experiments on the MIMIC-IV dataset demonstrated the feasibility of our proposed node pre-initialization technique.

In Paper C, we investigated diagnosis prediction from the perspective of graph-based EHR representations. While graph representations can leverage the relational dependencies of EHR data, it does not integrate the critical temporal dependencies. Hence, in the next chapter, we investigate sequence representations to structure EHR data. By transforming tabular EHR data into sequences of medical events pertaining to individual patient hospitalizations, we can better leverage the data’s temporal dependencies.



## Chapter 4

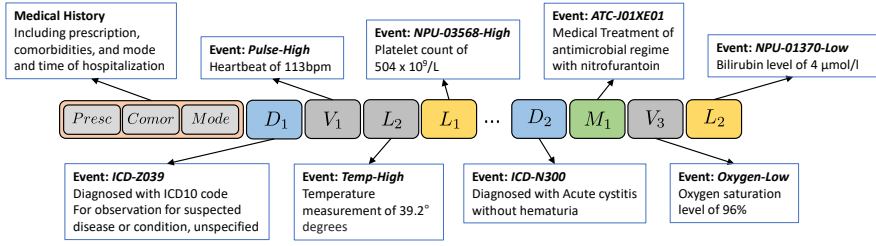
# Sequence-based EHR Representations

This chapter describes how we model EHR data as sequences of medical events for training a transformer encoder towards predicting the hospitalization time of patients. We investigate how measurement values can be encoded as part of the medical events using demographic-specific reference thresholds and domain taxonomies. The content in this chapter provides an overview of Paper D [18] and partly reuses its content.

### 1 Problem Motivation and Statement

In Chapter 3 we investigate the transformation of EHR data into graph representations in order to better integrate the relational dependencies of the data. However, graph representations neglect the temporal dependencies of the data. Hence, in this Chapter we investigate sequence representations of EHR data. We argue that EHR data inherently is sequential, with each patient’s medical history being a temporal sequence of events. Transforming tabular EHR data into sequence representations could allow us to capture the temporal dependencies and patterns over time in a patient’s healthcare data. However, transforming tabular data into sequences is not trivial. Hence, in Paper D, we investigate how medical domain taxonomies and demographic-specific reference threshold values can be used to create a finite lexicon for describing patients as sequences of medical events. Subsequently, we employ the transformer encoder DL model for learning to predict the length of patient hospitalizations.

Models that can predict the duration of a patient’s stay, or length of stay (LOS), can benefit healthcare facilities in resource management, such as staff



**Fig. 4.1:** Illustration of patient hospitalization structured as a sequence of medical events pertaining to the patient [18].

allocation and pre-emptive freeing of hospital beds. Applications for automatic forecasting of discharge times could be integrated into planning systems to alleviate hospital ward oversaturation [45]. Nonetheless, LOS prediction is a significant challenge due to the complexity of EHR data.

Historically, LOS prediction has been made on tabular EHR data since many ML models, such as the feedforward neural network, gradient boosting (GB), and support vector machines (SVM), already require the tabular data format as input. However, medical tabular EHR data often contain missing patient data values. Hence, imputation methods are often used to fill in the missing gaps [2]. However, missing observations in EHR data often are not missing at random (NMAR), which indicates that the absence of an observation itself carries significant information [28]. Furthermore, the tabular data format does not include temporal dependencies between observations, such as the chronological order between patient treatment events. To address these problems, we convert EHR data from a tabular structure into a sequential format, subsequently employing advanced DL technologies, we uncover novel prospects for deriving insights from the constantly growing corpora of healthcare data. Our approach can pave the way for enhanced decision support, ultimately leading to better quality of patient care.

The remainder of this chapter summarize the contributions of Paper D:

- First, we investigate the transformation of tabular EHR data into sequences of patient events.
- Second, We design a method of including measurement values into sequence events by utilizing domain taxonomies and demographic-specific reference thresholds.
- Third, we present the findings from Paper C, including the performance of using a transformer model for LOS prediction based on patient sequences.



## 2 Patient Sequence Representations

Patient hospitalizations can be naturally represented as sequences of medical events for evaluating, assessing, or diagnosing the patient’s health status. Healthcare facilities codify these events using standardized taxonomies such as the Anatomical Therapeutic Classification (ATC) [39] for medication events and ICD [7] for diagnosis events. Thus, a patient’s stay at the hospital can be expressed as a sequence of standardized concept tokens that capture the medical events pertaining to the patient. An example patient sequence is illustrated in Figure 4.1. As a patient’s medical history is vital to understand what treatment should be given, we incorporate the patient’s medical history as a tokenized vector pre-pended to the hospital sequence. At the beginning of the hospitalization, the patient is diagnosed with the ICD-10 code *Z039*. Subsequently, vital measurements and laboratory tests are conducted to monitor the patient’s health status and diagnose the underlying condition. As a result, the patient is diagnosed with acute cystitis without hematuria, denoted by the ICD-10 code *N300*, and the antibiotics nitrofurantoin (ATC code *J01XE01*) is prescribed. Following additional medical procedures and treatments, the patient is eventually discharged from the hospital. Hence, it is natural to structure a patient hospitalization as a sequence of medical events.

## 3 Event Measurements and Event Groupings

Numerical values often follow medical procedures and treatments like vital measurements and laboratory tests. Instead of disregarding these, we incorporate this data into patient sequences due to the crucial information they provide about a patient’s condition. For instance, knowing that a temperature reading was taken is in itself essential knowledge, but learning the measurement was  $40.1^{\circ}\text{C}$  indicates the patient has a fever. Using patient-specific thresholds for measurement values based on age, gender, and pregnancy status, we convert measurement values into tokens indicating normal, low abnormal, or high abnormal results. For example, if an albumin level of 56 g/L is measured for a 31-year-old male patient, we would create the *albumin-high* token to signify the measurement exceeded the expected range (36 – 48 g/L) for a patient with this demographic.

Furthermore, some measurement events are grouped together due to the nature of patient care and hospital administration. Nurses often perform several patient measurements, such as blood pressure, temperature, and heart rate, sequentially before updating the patient’s EHR. Consequently, we are often prevented from knowing the exact times and hence ordering of medical events. This issue is especially prevalent with laboratory tests, as multiple

## Chapter 4. Sequence-based EHR Representations

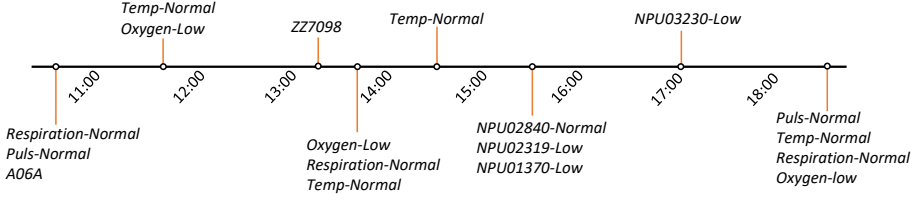


Fig. 4.2: Illustration of patient events grouping together [18].

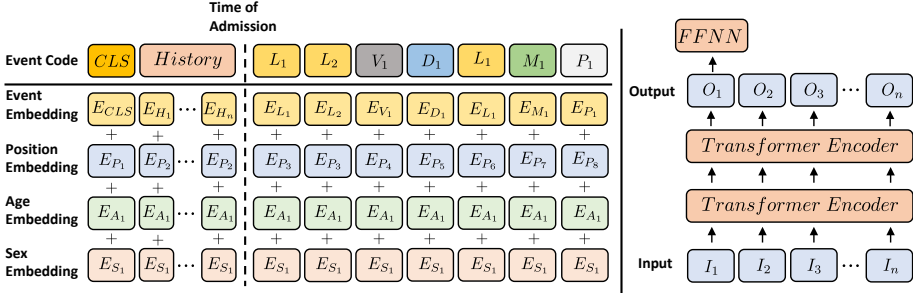


Fig. 4.3: Illustration of a patient event sequence embedding [18].

tests are often run on a single patient sample, making it impossible to determine the chronological order of event results. This problem is illustrated in Figure 4.2. We overcome this problem by assigning the same position embedding to events with the same start time, to ensure the model recognizes that these events do not have a fixed ordering. Using our specialized position embedding, we are also able to express the co-existence of complex medical events consisting multiple medical modalities. Figure 4.3 illustrates an event group by the position embedding where event codes  $L_1$  and  $L_2$  are mapped to the same position embedding  $E_{P_3}$ .

Additionally, just as it is common in medical reports to immediately state the age and sex of the patient, as these are essential factors for the treatment of the patient, so too do we put extra weight on these concepts by adding the age and sex embedding to every event embedding within a sequence. This can also be seen as a way of adding domain knowledge to the DL model [29].

## 4 Evaluation and Discussion

To investigate patient sequences for LOS prediction, we compiled a large Danish dataset of more than 48k emergency care patients from a large Danish hospital from 2018-2021. Using standardized medical taxonomies and demographic-specific threshold values, we transformed the tabular EHR data

## 5. Conclusion

into sequences of events pertaining to the patient hospitalizations. Furthermore, we created a modified version of the bidirectional encoder representations from transformers (BERT) [11] model, termed Medic-BERT (M-BERT) DL model, that can accommodate medical event groupings.

We compared our sequence-based approach to three standard tabular LOS prediction approaches: feedforward neural networks, random forest classification, and support vector machines. While random forest classification can naturally learn from missing data, the feedforward neural networks and support vector machines rely on imputation techniques for missing values. Experimental results from a binary, ternary, and regression problem demonstrated the feasibility of our approach.

These findings underscore the potential of employing transformer encoders for sequences of medical events. However, an interesting area of exploration lies in augmenting the method to more effectively assimilate and interpret the irregular intervals characterizing patient observations. At present, transformer encoders are constrained in that they predominantly focus on the ordering of events, neglecting the temporal durations between individual medical events.

## 5 Conclusion

In Paper D, we investigated how tabular EHR data can be transformed to sequence representations to integrate the temporal dependencies in the data. Furthermore, we could integrate the measurement values of medical events into the sequences using demographic-specific threshold values. Subsequently, we created and trained a specialized transformer encoder model, M-BERT, to accommodate the unique nature of EHR data, such as event groupings. Experiments on a large Danish dataset demonstrated the feasibility of our proposed method for LOS prediction and regression problems.

From the work of Paper C and Paper D, we have investigated graph representations and sequence-based EHR representations. While graph and sequence representations enable us to integrate disparate EHR modalities into the same data representation, none of them can naturally integrate and learn from raw medical modalities such as images, genome, and time-series data. Hence, in the next chapter, we investigate multi-modal representation learning (MRL) for medical modalities, which is the area concerned with learning from multiple modalities in an automated manner.



## Chapter 5

# Multi-modal Representation Learning

This chapter describes how we create a hierarchical taxonomy of medical information modalities and a hierarchical taxonomy of multi-modal representation learning technologies as a framework for surveying multi-modal representation learning techniques for medical applications. The content in this chapter provides an overview of Paper E [20] and partly reuses its content.

### 1 Motivation and Problem Statement

A patient's medical history can only be fully understood and represented using various medical observations, such as medical imaging events, microbiology events, clinical text, genomic sequences, etc., known as medical modalities. In the medical field, the progression of various diseases can be understood by observable changes in specific biomarker modalities, such as blood pressure, heart rate, and X-ray results. For instance, the progression of Alzheimer's Disease has been correlated with modalities like Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and protein measures of Cerebrospinal Fluid (CSF) [5, 8, 27]. Each of these modalities offers distinct insights that, when combined, could provide added insights into complex medical tasks such as predicting Alzheimer's progression.

While machine learning aims to enhance the quality and speed of traditional manual medical tasks, many tasks are still predominantly done using single-modality approaches. However, integrating multiple complementary medical modalities into machine learning models could boost the performance of predictive models.

Multi-modal representation learning (MRL) is the field concerned with combining diverse medical modalities to enhance machine learning tasks [3]. Recently, MRL has made its way into the medical domain, where it has been used to combine various medical modalities for diagnosis and prognosis tasks. However, no comprehensive survey has been conducted on the application of MRL in the healthcare domain. Numerous medical modalities exist in healthcare, including genomics data, medical images, textual medical records, electronic health records (EHR), clinical practice guidelines, and biomedical knowledge graphs. Navigating the vast number of distinct modalities, the prospect of their combination for enhanced medical tasks, and the technologies that can be used to combine them is challenging.

In Paper E, We investigate the landscape of MRL from the perspective of the medical domain. Starting with structured and unstructured modalities, we build a three-level hierarchical taxonomy for organizing medical information modalities. The modality taxonomy is one of two taxonomies we create for structuring related work in MRL in the medical domain. The second taxonomy structures the techniques of MRL into three main classes, namely Alignment, Fusion, and Neural techniques, with two further hierarchical levels for granular classification of individual approaches.

The remainder of this chapter summarize the contributions of Paper E.

- First, we create a hierarchical taxonomy for organizing medical information modalities into three levels of increasing granularity.
- Second, we build a hierarchical taxonomy of MRL techniques, with Neural deep learning techniques as one of the main classifications.
- Third, we make a comprehensive literature survey of more than 1000 papers on combining medical modalities for solving medical analytics tasks.

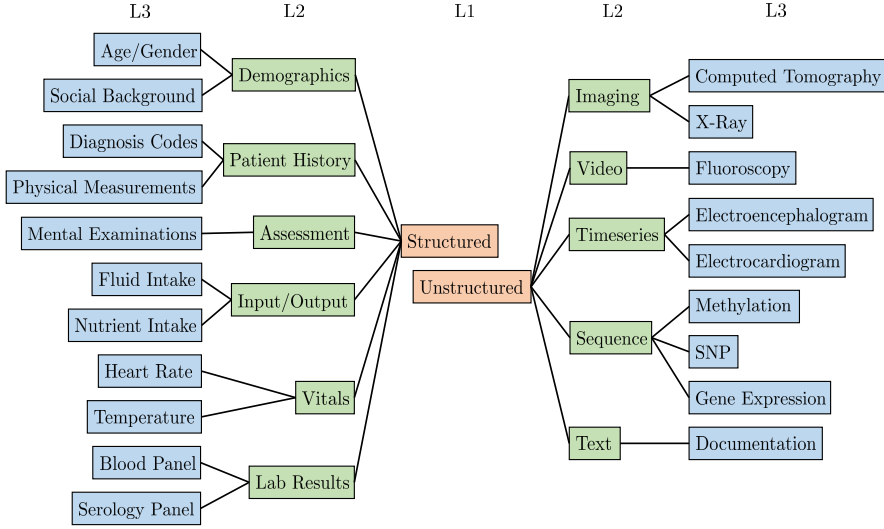
## 2 Medical Information Modality Taxonomy

Medical information modalities describe patients in structured and unstructured ways to get insights into the patient’s current and future healthcare status. Hence, medical information modalities are the prime source of information for medical analytics systems.

Structured data is an organization of distinct measurements containing specific values, such as a relational database where a patient’s health status can be described using specific predefined attributes in a fixed way. In medicine, structured data modalities could be vital measurements, laboratory results, and demographic data.

On the other hand, unstructured data is less straightforward to understand. It contains many data points that do not hold any meaning when viewed individually but collectively, if interpreted correctly, provide insights

## 2. Medical Information Modality Taxonomy



**Fig. 5.1:** Partial hierarchy for structuring medical modalities. The full hierarchy can be found at <https://tabsoft.co/40aAECd> [20].

into the patient’s health status. Take the X-ray image as an example. Individual pixel values of the image do not provide any information, but by looking at the whole image, a trained professional might identify a fracture in the tibia. Besides imaging modalities, other forms of unstructured data include video, time series, genomic sequences, and text.

Numerous medical taxonomies exist for organizing medical concepts into categories and groups with terminologies such as ICD-10, SNOMED, and ATC. However, currently, no comprehensive taxonomy for organizing medical modalities exist. In Paper E, we propose a three-level hierarchical taxonomy for organizing medical modalities.

Figure 5.1 depicts part of our proposed hierarchy. The top-level structures modalities as either structured or unstructured data. The second level categorizes modalities in groups familiar in machine learning literature, such as images, text, and timeseries data. The third level specifies specific medical information modalities used to get insights into the healthcare status of patients. Furthermore, we connect the third-level modalities to SNOMED taxonomy concepts to enable smooth linkage to other terminologies and taxonomies. For example, the level three concept *Computed Tomography* aligns with the SNOMED concept *Computed tomography (procedure)*. The SNOMED concept can then be mapped to other taxonomies such as MedDRA’s *CT scan* and BIM’s *Computed\_tomography* concepts. The entire hierarchy can be accessed online.

Thus, our medical modality hierarchy provides an excellent framework

for studying MRL approaches in the medical domain.

### 3 Multi-modal Representation Learning Taxonomy

It is a common challenge in medical analytics to deal with multiple medical modalities. Different data types, such as medical images, genomics, and clinical data, provide complementary perspectives on a patient's health status. However, integrating such diverse modalities in a meaningful way is a complex task. The first step in investigating what MRL technologies could help combine specific medical analytics is knowing what technologies exist and how they have been used. Hence, we create a hierarchical taxonomy for MRL methods that can be used to structure the technologies used in the amalgamation of modalities for medical analytics.

The taxonomy can serve as an entry point for researchers, practitioners, and newcomers to the field of MRL, enabling them to quickly get an overview of the landscape of approaches to find the one that best suits their needs. Furthermore, it can inspire new ideas by highlighting novel MRL techniques within the medical domain.

We break down MRL techniques into Alignment, Fusion, and Neural approaches as illustrated in Figure 5.2. Alignment methods aim to identify a feature space where modalities can co-exist, fusion methods merge uni-modal features into a shared representation space, and neural methods jointly learn a latent representation that combines uni-modalities for solving a specific medical analytics task.

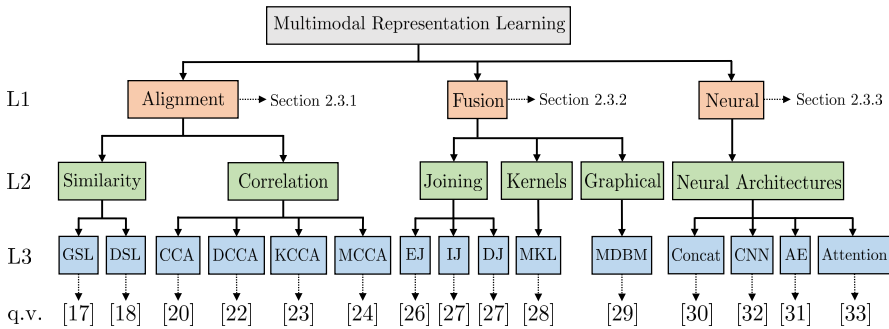


Fig. 5.2: Taxonomic hierarchy of reviewed MRL techniques used in medical analytics [20].

Alignment-based MRL finds a representation space where uni-modal modalities  $x$  and  $y$  can co-exist. This group of techniques works on the premise that similar samples should be closer to each other in the learned space than dissimilar ones. Mathematically this can be expressed as  $f(x_i) \sim$



$g(y_i)$  where  $g$  and  $f$  are modality-specific projection functions that map individual samples  $x_i$  and  $y_i$  into a multi-modal space aligned by the distance function  $\sim$ . We further categorize alignment-based techniques into correlation and similarity-based techniques.

Fusion-based MRL can be mathematically defined as  $z = \phi(x_i, y_i)$ , where  $\phi$  is a function that combines uni-modal modality samples  $x_i$  and  $y_i$  into a merged latent representation  $z$ . Fusion techniques are typically used to boost the performance of medical analytics models in cases where distinct uni-modalities possess unique discriminative properties [51]. We divide fusion-based MRL techniques into joining, kernels, and graphical techniques with complexities ranging from straightforward feature concatenation to intricate kernel combinations.

Neural architectures aim to combine uni-modal representations through supervised, semi-supervised, and unsupervised methods. They share the idea of utilizing layers of non-linear trainable transformations to fuse uni-modalities into a latent representations space, guided by optimizing a loss function that targets a specific medical analytic [12].

Our MRL taxonomy enables us to evaluate MLR techniques in the context of combining uni-modal medical modalities. As the best technique for the amalgamation of uni-modal medical modalities will always depend on the data and the medical analytics, our taxonomy provides a comprehensive, structured overview of the MRL landscape that can be explored for insights and ideas in building multi-modal medical analytics.

## 4 Literature Survey of MRL for Medical Analytics

A last contribution of this work is the systematic literature survey we conducted in multi-modal representation learning for medical analytics using our constructed hierarchical MRL and medical modality taxonomies. We surveyed more than 1000 scientific articles in the PubMed search engine using the PRISMA flow chart for reporting systematic reviews [34]. The resulting classification facilitated the exploration of each paper’s contribution to the field within the scope of the taxonomies. Furthermore, the classification made it possible to discover patterns, such as the number of papers using neural technologies to combine various medical modalities.

Moreover, we provide an interactive online analysis as an electronic supplement accompanying this work to provide researchers with a platform for investigating the literature survey themselves. Using the platform, we elucidate intriguing aspects of the literature survey, such as the commonality of modality pairings and the prevalence of MRL techniques across medical analytics. The frequency of modality combinations indicates that unstructured brain imaging modalities are often combined with other unstructured

brain imaging modalities for medical analytics related to the nervous system. Moreover, there is a lack of research combining structured data with unstructured modalities. The lack could be a potential area for future research. Furthermore, the survey shows a substantial prevalence of neural joining techniques for amalgamating structured modalities such as laboratory results and demographics data.

The review further demonstrates that most studies focus on analytics-targeted nervous system diseases. We believe this finding originates from the popularity of open datasets related to investigating Alzheimer’s disease, such as ADNI [35]. As diseases of the circulatory system are a significant focus area for medical AI research, the lack of MRL research points towards the need for more openly available datasets in this field. Furthermore, we observe that most medical analytics are predictive.

This research contributes to creating a taxonomy for medical modalities, MRL techniques, and a comprehensive literature survey into multi-modal representation learning for medical analytics. While progress has been made in applying MRL to medical analytics, there remains vast untapped potential, especially in less-studied disease categories such as diseases related to the circulatory system. For research to flourish, an obstacle we must overcome is the collective effort to make medical data repositories openly available for researchers to continue investigating new ideas in the amalgamation of modalities for creating medical analytics.

## 5 Conclusion

In Paper E, we investigate the field of MRL from the perspective of medical analytics. To study MRL from the medical domain, we first create a novel three-layer hierarchical taxonomy for structuring medical modalities. The first level classifies modalities into structured and unstructured modalities. The second level groups modalities based on classical groups from the machine learning literature. The third level specifies specific medical information modalities. Furthermore, we create a hierarchical taxonomy of MRL technologies for merging and aligning disparate modalities.

Based on the MRL and modality hierarchies, we created a systematic literature survey of more than 1000 scientific articles to discover what MRL technologies have been previously used to combine specific medical modalities. Furthermore, we made the literature survey publicly available as an interactive online analysis for researchers to investigate the survey for themselves.

## Chapter 6

# Conclusions and Future Work

Deep learning for EHR data is becoming a large and important field of research due to its prospects for the future of the healthcare domain. DL technologies for the healthcare domain hold the prospects to complement the work of medical professionals by providing timely and accurate decision support for improved diagnostic accuracy in decision-making, reducing the time to diagnosis, and providing more personalized patient care. In this thesis, we investigate complex aspects of deep learning for EHR data, such as integrating medical domain knowledge and transforming tabular EHR data into representations such as sequences and graphs. Advances in these areas are essential to realize the full potential of DL for EHR data and ultimately provide the best patient care.

In summary, each research paper provides the following contribution:

- In Paper A [41], we examine the novel task of diagnosis prediction based on a patient’s medication history. We develop a language-agnostic DL-based system that can accurately predict the diagnosis codes of a patient, given a patient’s medication list. We developed a novel loss function to advance the system’s performance to distinguish minor from significant model errors using the similarity between diseases from the hierarchical ICD taxonomy.
- In Paper B [22], we investigate the transferability and generalizability of the system and novel loss function developed in Paper A by using and testing our approach on a large Danish dataset consisting of more than 2 million patients. To facilitate interoperability between the datasets, we created a mapping between diagnosis concepts from the Danish ICD-10-SKS and the American ICD-9-CM taxonomies and between the ATC and the RxNorm vocabularies using the OMOP CDM.
- In Paper C [19], we transform tabular EHR data into graph representations for subsequent patient diagnosis prediction. Graphs elegantly

overcome the prevalent problem of missing data and enable learning from the relational dependencies of EHR data. We transform tabular patient EHR data into graphs consisting of nodes and edges connecting patients with their clinical observations. We can learn from the graph’s structure by leveraging the graph convolution neural network technology. Furthermore, we investigate a novel node pre-initialization method based on extracting the structure of medical hierarchical taxonomies. Such embedding enables the model to work in an inductive setting by keeping node embeddings stable during training.

- In Paper D [18], develop a method for transforming tabular EHR data into sequences representing the medical events of patient hospitalizations and leverage a specialized transformer encoder DL technology to learn from the temporal order of patient events. Furthermore, we were able to integrate event measurement values into the event concepts by leveraging demographic-specific threshold values. To further advance the system’s performance, we investigate a position encoding that enables groups of sequence events to be unordered.
- In Paper E [20], we investigate the field of multi-modal representation learning from the perspective of medical analytics. We build two novel hierarchical taxonomies, one for structuring medical modalities and one for structuring multi-modal representation learning techniques. The taxonomies allow us to conduct a structured literature survey of more than 1000 scientific papers for investigating previous work in MRL for medical analytics. Furthermore, we made the survey publicly available as an interactive online analysis.

## Future Work

The papers outline multiple paths of future work. However, the following three are vital for DL technologies to be widely used in a clinical setting.

Firstly, to fully harvest the benefits of DL for EHR data, we need to improve our integration of diverse medical domain knowledge with DL technologies. In our work, we have been focusing on hierarchical medical domain taxonomies; however, this is only one type of knowledge that could be used to improve the performance of DL systems in the medical domain. Biomedical knowledge, such as our understanding of disease pathways, could be used in the model design or to create more sophisticated features. Understanding disease prevalence and risk factors, such as prior knowledge about the prevalence of certain diseases, could be used to improve the model performance for imbalanced datasets. Furthermore, integrating drug-drug interactions, drug-disease interactions, and pharmacokinetics into DL models is an exciting research direction. These are only a few of the possible types of domain

knowledge that, if added to DL models, could improve their performance.

Secondly, the graph and sequence representations we have investigated in this thesis are great representations for expressing EHR data. Furthermore, transformer encoders and graph convolution neural networks are compelling DL models for learning from these representations. However, DL models are not inherently explainable. In order to use DL systems in a clinical context, we need an explanation for their predictions. It is not enough to tell that a patient will likely have a prolonged hospitalization; we want to know the reason for this prediction. As the decisions of medical decision systems can have a profound impact on the treatment of patients, clinicians need the models to explain their predictions to facilitate trust in the DL systems. Moreover, the European Union’s General Data Protection Regulation (GDPR) regulatory requirements include a clause on the patient’s right to explanation. Thus, medical decision support systems are required to be able to explain their decisions. Lastly, DL systems are often prone to learn from biases in medical data, often due to imbalances in our medical datasets. Hence, explainability methods can facilitate finding and correcting such biases.

Thirdly, EHR data is intrinsically temporal in nature. Medical events of different types exhibit varying temporal granularities with measurement frequencies ranging from seconds to years. Additionally, medical events can occur simultaneously and relate to each other using the spectrum of Allen’s 13 temporal relations. The deep learning models used in this thesis predominantly overlook the integration or learning from the temporal dimensions inherent to EHR data, with the notable exception of the transformer encoder that learns from the sequential ordering of medical events. There lies significant potential in exploring the augmentation of deep learning technologies, such as the graph convolution neural network or the transformer encoder to better harness and learn from the temporal aspects of EHR data.

## References

- [1] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield, "Efficient graphlet counting for large networks," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 1–10.
- [2] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar, "Machine learning in the prediction of medical inpatient length of stay," *Intern. Med. J.*, vol. 52, no. 2, pp. 176–185, 2022.
- [3] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [4] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Graph neural networks in network neuroscience," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5833–5848, 2022.
- [5] K. Blennow, "Cerebrospinal fluid protein biomarkers for alzheimer's disease," *NeuroRx*, vol. 1, no. 2, pp. 213–225, 2004.
- [6] J. D. Bossér, E. Sörstadius, and M. H. Chehreghani, "Model-centric and data-centric aspects of active learning for deep neural networks," in *IEEE BigData*. IEEE, 2021, pp. 5053–5062.
- [7] D. J. Cartwright, "Icd-9-cm to icd-10-cm codes: what? why? how?" 2013.
- [8] R. E. Coleman, "Positron emission tomography diagnosis of alzheimer's disease," *PET Clinics*, vol. 2, no. 1, pp. 25–34, 2007.
- [9] C. R. Cooke, M. J. Joo, S. M. Anderson, T. A. Lee, E. M. Udris, E. Johnson, and D. H. Au, "The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease," *BMC health services research*, vol. 11, no. 1, p. 37, 2011.
- [10] E.-M. Dalsgaard, D. R. Witte, M. Charles, M. E. Jørgensen, T. Lauritzen, and A. Sandbæk, "Validity of Danish register diagnoses of myocardial infarction and stroke against experts in people with screen-detected diabetes." *BMC public health*, vol. 19, no. 1, p. 228, feb 2019.
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT '19*, 2019, pp. 4171–4186.
- [12] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
- [13] N. Garg, M. S. Choudhry, and R. M. Bodade, "A review on alzheimer's disease classification from normal controls and mild cognitive impairment using structural mr images," *Journal of neuroscience methods*, vol. 384, p. 109745, 2023.
- [14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

## References

- [15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [16] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.
- [17] —, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.
- [18] E. R. Hansen, T. D. Nielsen, T. Mulvad, M. N. Strausholm, T. Sagi, and K. Hose., "Hospitalization length of stay prediction using patient event sequences," *Artificial Intelligence in Medicine*, 2023.
- [19] E. R. Hansen, T. Sagi, and K. Hose, "Diagnosis prediction over patient data using hierarchical medical taxonomies," in *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference*, vol. 3379. CEUR-WS.org, 2023. [Online]. Available: [https://ceur-ws.org/Vol-3379/HeDAI\\_2023\\_paper400.pdf](https://ceur-ws.org/Vol-3379/HeDAI_2023_paper400.pdf)
- [20] —, "Multi-modal representation learning for medical analytics," in *Unpublished manuscript*, 2023.
- [21] E. R. Hansen, T. Sagi, K. Hose, G. Y. H. Lip, T. B. Larsen, and F. Skjøth, "MIMIC Prescriptions result files," 2020.
- [22] E. R. Hansen, T. Sagi, K. Hose, G. Y. Lip, T. B. Larsen, and F. Skjøth, "Assigning diagnosis codes using medication history," *Artificial Intelligence in Medicine*, vol. 128, p. 102307, 2022.
- [23] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics*, vol. 216, pp. 574–8, 2015.
- [24] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes," *Computer methods and programs in biomedicine*, vol. 177, pp. 141–153, 2019.
- [25] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.
- [26] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, may 2016.
- [27] R. J. Killiany, T. Gomez-Isla, M. Moss, R. Kikinis, T. Sandor, F. Jolesz, R. Tanzi, K. Jones, B. T. Hyman, and M. S. Albert, "Use of structural magnetic resonance imaging to predict who will get alzheimer's disease," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 47, no. 4, pp. 430–439, 2000.

## References

- [28] J. Li, X. S. Yan, D. Chaudhary, V. Avula, S. Mudiganti, H. Husby, and et al., "Imputation of missing values for electronic health record laboratory data," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–14, 2021.
- [29] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "Behrt: transformer for electronic health records," *Scientific reports*, vol. 10, no. 1, p. 7155, 2020.
- [30] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "Rxnorm: prescription for electronic drug information exchange," *IT professional*, vol. 7, no. 5, pp. 17–23, 2005.
- [31] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for health-care: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.
- [32] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, pp. 1207–1220, 2021.
- [33] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical Natural Language Processing in languages other than English: Opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, no. 1, mar 2018.
- [34] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105906, 2021.
- [35] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga *et al.*, "Alzheimer's disease neuroimaging initiative (adni): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.
- [36] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 938–947, 2015.
- [37] F. Pontet, U. M. Petersen, X. Fuentes-Arderiu, G. Nordin, I. Bruunshuus, J. Ihalaainen, and et al., "Clinical laboratory sciences data transmission: the npu coding system," *Stud. Health Technol. Inform.*, vol. 150, p. 265, 2009.
- [38] N. Razavian, J. Marcus, and D. A. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, ser. JMLR Workshop and Conference Proceedings, F. Doshi-Velez, J. Fackler, D. C. Kale, B. C. Wallace, and J. Wiens, Eds., vol. 56. JMLR.org, 2016, pp. 73–100. [Online]. Available: <http://proceedings.mlr.press/v56/Razavian16.html>
- [39] M. Ronning, "A historical overview of the atc/DDD methodology," *WHO drug information*, vol. 16, no. 3, p. 233, 2002.
- [40] R. A. Rossi, R. Zhou, and N. K. Ahmed, "Deep inductive network representation learning," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 953–960.



## References

- [41] T. Sagi, E. R. Hansen, K. Hose, G. Y. Lip, T. Bjerregaard Larsen, and F. Skjøth, "Towards assigning diagnosis codes using medication history," in *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*. Springer, 2020, pp. 203–213.
- [42] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, 2021.
- [43] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.
- [44] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.
- [45] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework," *PLOS Digital Health*, vol. 1, no. 4, p. e0000017, 2022.
- [46] F. Teng, Z. Ma, J. Chen, M. Xiao, and L. Huang, "Automatic medical code assignment via deep learning approach for intelligent healthcare," *IEEE journal of biomedical and health informatics*, vol. 24, no. 9, pp. 2506–2515, 2020.
- [47] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, jan 2018.
- [48] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical Multi-Label Classification Networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 5075–5084. [Online]. Available: <http://proceedings.mlr.press/v80/wehrmann18a.html>
- [49] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.
- [50] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligence magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [51] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative *et al.*, "Multimodal classification of alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.
- [52] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [53] S. Zhang, S. M. H. Bamakan, Q. Qu, and S. Li, "Learning for personalized medicine: a comprehensive review from a deep learning perspective," *IEEE reviews in biomedical engineering*, vol. 12, pp. 194–208, 2018.

## References

- [54] Z. Zhao, H. Zhou, L. Qi, L. Chang, and M. Zhou, "Inductive representation learning via cnn for partially-unseen attributed networks," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 695–706, 2021.

# **Part II**

# **Papers**



# Paper A

## Towards Assigning Diagnosis Codes Using Medication History

Tomer Sagi, Emil Riis Hansen, Katja Hose, Gregory Y. H. Lip,  
Torben Bjerregaard Larsen, Flemming Skjøth

The paper has been published in the  
*Proceedings of the 18th International conference on Artificial Intelligence in  
Medicine (AIME 2020)*, pp. 203-213, 2020.

DOI: [10.1007/978-3-030-59137-3\\_19](https://doi.org/10.1007/978-3-030-59137-3_19)

## Abstract

*Prior studies have manually assessed diagnosis codes and found them to be erroneous/incomplete between 4–30% of the time. Previous methods to validate and suggest missing codes from medical notes are limited in the absence of these, or when the notes are not written in English. In this work, we propose using patients' medication data to suggest and validate diagnosis codes. Previous attempts to assign codes using medication data have focused on a single condition. We present a proof-of-concept study using MIMIC-III prescription data to train a machine-learning-based model to predict a large collection of diagnosis codes assigned on four levels of aggregation of the ICD-9 hierarchy. The model is able to correctly recall 58.2% of the ICD-9 categories and is precise in 78.3% of the cases. We evaluate the model's performance on more detailed ICD-9 levels and examine which codes and code groups can be accurately assigned using medication data. We suggest a specialized loss function designed to utilize ICD-9's natural hierarchical nature. It performs consistently better than the non-hierarchical state-of-the-art.*

© Springer Nature Switzerland AG 2020. Reprinted, with permission, from Tomer Sagi, Emil Riis Hansen, Katja Hose, Gregory Y. H. Lip, Torben Bjerregaard Larsen and Flemming Skjøth.

Towards Assigning Diagnosis Codes Using Medication History. In: Proceedings of the 18th International conference on Artificial Intelligence in Medicine (AIME 2020), Lecture Notes in Computer Science, volume 12299, pages 203–213, September, 2020. [https://doi.org/10.1007/978-3-030-59137-3\\_19](https://doi.org/10.1007/978-3-030-59137-3_19)

*The layout has been revised.*

# 1 Introduction

The practice of coding diagnoses of medical conditions using standardized coding systems such as ICD-10 [25] has grown prevalent. However, while coding systems are in wide-spread use, coding quality is uneven. Coding a medical diagnosis is notoriously complex. There exist multiple hierarchies and choosing the appropriate code requires a deep understanding of their structure and the relationships. For example, in a review of 1800 injury discharges from a New Zealand hospital, Davie et al. [6] found 2% to be uncoded, and 14% of PDx codes and 26% of external cause codes to be inaccurately coded. Wockenfuss et al. [26] determined that ICD-10 three and four level codes are too detailed to be reliable for general practitioners by measuring the Kappa inter-rater agreement scores. and found a sensitivity (recall) of 93.4 and positive predictive value (precision) of 88.9. Some work exists on predicting diagnoses from laboratory results (e.g., [19]), but is limited to cases where such results are available and relevant. A large body of work exists on extracting diagnoses from clinical notes and reports (see review [23]). However, these systems' performance is reliant on techniques that tend to work much better in English, and must be retrained for every new language [17].

A patient's current medication can shed valuable light on their existing medical conditions. For example, observing that a patient has a chronic prescription for Metoprolol usually indicates that he/she is suffering from hypertension or ischaemic heart disease. Generalizing upon this observation, in this work we develop a machine-learning-based model able to predict the list of diagnoses assigned to a patient based upon his/her medications. Furthermore, in some countries (e.g., Denmark [20] and South Korea [15]) centralized medication repositories are comprehensive, while diagnosis codes are sporadic. Thus, such a model could provide emergency responders and critical care facilities with a rapid assessment of a patient's existing conditions in addition to the model's utility in diagnoses quality control. For example, an unconscious patient with a history of diabetes, will be first assessed for hyper/hypoglycemia, while one without a history of diabetes, but with a history of heart-disease, will be first assessed for acute heart conditions such as a heart-attack. We assess the viability of our approach using the publicly available MIMIC-III dataset [14]. The dataset contains rigorously anonymized and detailed medical records for over 50K ICU patients.

## 2 Related Work

The need to perform quality control of diagnosis code assignment is justified by several studies. Cooke et al. [4] have shown that an ICD-9 code as a

predictor of true COPD had a sensitivity of 76% and specificity of 67% using spirometry as their golden standard. A validity study of Danish national registry diagnoses [5] showed that only 75% of diabetic patients labeled with MI or stroke actually had such an event. Recent work attempted to predict ICD-9 assignment in MIMIC-III from discharge notes [12]. Their solution to the problem of multi-label, multi-level was to either limit the number of labels or aggregate predicted codes into categories, thereby solving two separate problems, namely to predict the top-10/50 codes or the top 10/50 categories. In this work, we aim to predict all codes, at different aggregation levels, in order to examine which codes and code groups can be predicted from medication data.

There have been a few attempts to use prescription data to predict a single or at most two conditions. Schmidt et. al. developed and validated an algorithm with 87% accuracy able to identify herpes zoster [22]. In another study, prescription data was used to classify whether or not patients had preexisting conditions of diabetes or hypertension [21]. In a recent review [8] of algorithms designed to extract cases for medical research from EMR data, some of the studies use medication data. However all studies extract cases for a single condition, often aggregating several diagnosis codes. In our scenario, we identify the probable diagnosis codes of multiple conditions at once and thus identify cases where improbable diagnosis codes have been used.

### 3 Methods and Data

#### 3.1 Data

We use MIMIC-III [14] from PhysioNet [9], EHR data for 50K patients from an American hospital’s ICU departments over four years. MIMIC-III contains an extensive variety of data, including lab results, vital signs, medical notes, and most importantly for our needs, drugs administered and diagnoses ascertained. The prescriptions table (model input) contains 4M rows of drugs prescribed during 50,216 admissions. There are 4,525 different drug names in the DRUG field, which are often the same drug, with different spelling or with an added comment, e.g., *Basiliximab* and *\*NF\* Basiliximab*. To disambiguate and standardize the codes we use a mapping of MIMIC terms to the OMOP concepts [11] and group them by *Clinical Drug Form* to receive 1,602 RxNorm drug codes.

The diagnosis table (expected output) contains 651,048 diagnoses for 58,925 admissions using 6,841 different ICD-9 codes. ICD-9 is a hierarchical grouping of disease codes that consists of 5 levels starting from 0 (most general), to 4 (most specific). ICD-9 is built on the basis of grouping for similar disease. Upon review, we omit 5,994 codes for which less than 100



cases exist as it is typically not possible to generalize from such a low number. We further omit a number of codes focusing on diagnoses for chronic or persistent conditions. A complete and more detailed description of omissions can be found in the appendix. We use the patient data to add the *age* in years upon admission. MIMIC hides elderly (over 89) patient ages due to anonymization concerns and reports an average of 92.4 for the patients over 89. We use this age as the replacement age for these patients and further normalize the age by dividing it by 92.4, a practice that has been shown to be beneficial in machine learning techniques. After joining with the prescriptions table, the final table contains 52K admissions of 40K different patients using 567 unique codes, denoted labels in the following.

## 3.2 Task - Hierarchical Multi-label Classification (HMC)

Binary classification problems (e.g., will this person develop Sepsis ) aim to correctly classify each task as either positive or negative. Single-label multi-class problems (e.g., is the following brain MRI normal, or does it contain a glioblastoma, a sarcoma, or a metastatic bronchogenic carcinoma?), extend the classification to allow more than one class for each task. These two types of ML tasks are, by far, the most commonly studied in the medical domain. Less common are multi-label classification problems which attempt to assign a set of labels to each example (e.g., which of the ICD-9 codes should be assigned following this medical report [1]), each of the labels is drawn from a possible set of classes. Since each person may have multiple co-morbidity, the task of assigning the correct set of diagnosis codes can be characterized as a multi-label classification problem [27]. The hierarchical nature of diagnoses both complicates the task and offers an opportunity to improve its applicability. If an algorithm predicts a patient suffering from non-specified chiroisis (ICD-9 code 571.5) to be suffering from alcoholic chiroisis (ICD-9 code 571.2) it should be more appreciated than if no chiroisis related diagnosis are returned since both codes share a common ancestor. Further hierarchical constraints may dictate that a person cannot have more than one label from the same sub-tree of codes. Since ICD-9 is indeed hierarchical and imposes such constraints on some of its sub-trees, we can classify our task as an hierarchical multi-label classification (HMC) problem.

## 3.3 Machine Learning and Loss Functions

Many approaches to HMC include splitting the problem into multiple simple (single label) classification tasks, each of which is trained separately. Within these approaches, local and global approaches [7] differ by the amount of classifiers trained. In the local case, multiple classifiers are trained over a binary label pertaining to a single node in the hierarchy and the predictions of

each level are subsequently propagated. In the global case, the labels are selected from a set of all possible labels. In this work we follow the observation of Cerri et al. [2] that by training a single global classifier based on a multi-level neural network representation, one can effectively reuse the high-level features learned to discriminate between high levels in the hierarchy and then refine these to more accurate code assignments using the subsequent levels of the neural network. Furthermore, deep neural networks (DNN) have repeatedly shown superiority over other techniques in the medical domain (e.g., [13], [3]). We therefore employ a multi-layer perceptron, or fully connected neural network. The input layer for this network is comprised of one node for each RxNorm code in the data (and one for normalized age) and the output layer of one node for each ICD-9 code at the chosen roll-up level. The number of internal layers and the number of nodes in each layer are hyper-parameters over which we perform a classic grid-search.

Machine learning, in particular deep learning, uses a loss function during the training phase to quantify the error of the current iteration of the model with respect to the expected output. Choosing an appropriate loss function is crucial and in general must reflect the structure of the expected output. Thus, specific loss functions have been suggested for the multi-label case [16] as well as hierarchical multi-label functions [24]. However, these are tied directly to the structure of the global classifier, and none have been applied in the medical data setting using the inherent hierarchy of a medical ontology.

We therefore experiment with two types of loss functions. One suitable for the multi-label case, where each missed label is treated the same regardless of the extent of the mistake, and one designed for the HMC case. Our multi-label function is the multi-label soft margin loss function [28], defined as follows with  $C$  being the number of classes  $y$  being the class indicator and  $x$  the current value of the corresponding output node ( $i$  iterates over all classes).

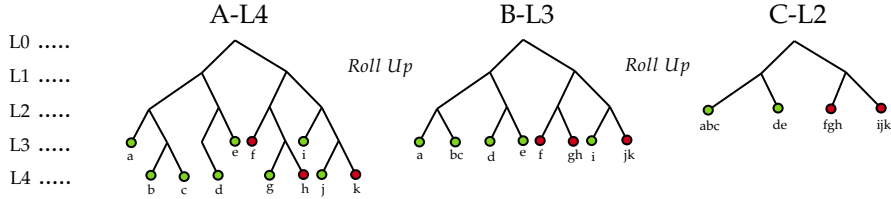
$$\begin{aligned} \text{loss}(x, y) = & -\frac{1}{C} \sum_i y[i] \cdot \log((1 + \exp(-x[i]))^{-1}) + \\ & (1 - y[i]) \cdot \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right) \end{aligned} \quad (\text{A.1})$$

We model our HMC loss function (hml, Eq. A.2) after the one developed for HMCN-F [24], while adjusting it to account for the differences between a text-classification problem and our own task and minimize a function comprised of two components.

$$\mathcal{L}_{hml} = \mathcal{L}_L + \mathcal{L}_G \quad (\text{A.2})$$

$\mathcal{L}_L$  is the local loss – calculation of Eq. A.1 at the leaf level.  $\mathcal{L}_G$  is calculated by rolling up the results one layer at a time until the ICD-9 chapter

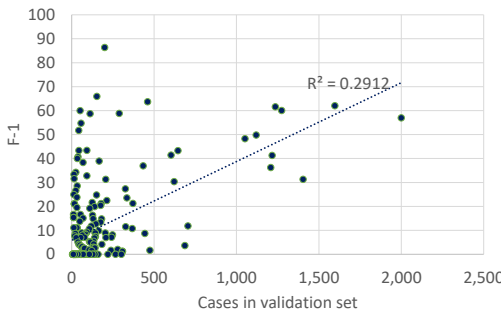
level (0). At each phase of the roll-up, the predictions for each inner node are set to the average of the predictions over its children. The loss of each level is calculated and summed to the other levels. Since our neural network does not directly predict the global scores, we do not suffer from hierarchical violations and do not require the third component that penalizes them in HMCN-F. We employ the Roll Up method to aggregate diagnoses given the



**Fig. A.1:** Example of the roll up algorithm. An example level 4 code assignment is shown as tree A-L4. Disease codes {b, c, d, g, f, j, k} are level 4 billable codes, whereas codes {a, e, f, i} are billable codes on level 3. Red circles are the registered comorbidities of the patient. Green circles are diseases not recorded in the patient.

ICD-9 hierarchy (see example in Figure A.1). A disease is only billable if it is a leaf-node of the ICD-9 hierarchy. However, not all leaves are on the same level. As an example, the code 322.2 is a billable level 4 code, which represents *Chronic meningitis*, whereas code 003.22 is a billable level 5 code for *Salmonella pneumonia*. Each patient initially starts with one or more billable disease from the ICD-9 hierarchy.

### 3.4 Evaluation



**Fig. A.2:** F-1 by number of cases over level 2 codes.

A few previous ICD coding tasks have been evaluated by measures that consider its hierarchical nature as well [18]. To allow comparison of ICD code assignment using medication data to algorithms using medical notes, we use the more common micro-averaged precision and recall, and their harmonic mean F1. We perform the experiments on different prediction resolutions.

With level 0 corresponding to the chapter level of ICD-9 (e.g., 520–579: diseases of the digestive system) and level 1 to the code group level (e.g., 401–405

Hypertensive Disease). Our last level corresponds to the most detailed available in the ICD-9 hierarchy (level 4) with 576 possible codes.

## 4 Results

Table A.1 presents the best results (by F1) obtained by the ML-model following a standard hyper-parameter grid search. In each task, the code assignments were rolled up prior to both the training and the test phase and not only for the purpose of evaluation, such that the neural network encountered a different task for each level. For each ICD level we provide the number of codes in that level, the average branching factor, and the average number of eventual leaves a node in this level’s sub-tree. In addition to precision, recall, and F1, we show the number of diagnosis codes for which F1 was equal to zero.

Since this is a relatively small dataset, the number of cases for many diagnoses is too low to expect reasonable performance. When examining the effect of the number of cases on the model’s performance (Fig. A.2) we find that at least some of the variance can be explained by the small number of cases ( $R^2$  of 0.29 for a linear model). Top-5/top-10 results by code are available as an online appendix containing the full results [10].

**Table A.1:** MIMIC-III Diagnosis Prediction Results

Prediction Task	Codes	Branch	Avg. Leaves	Prec	Rec	F1	F1=0
Top-10-groups (L0)	10	NA	NA	82.6	52.4	64.1	0
Top-10-codes (L4)	10	NA	NA	61.3	50.5	55.4	0
Rolled Up (L0)	15	5.7	565.1	78.3	58.2	66.8	2
Rolled Up (L1)	65	8.4	108.3	60.9	43.0	50.4	15
Rolled Up (L2)	236	6.6	14.0	56.6	31.5	40.5	122
Rolled Up (L3)	461	1.6	1.6	52.3	19.9	28.8	297
Raw (L4)	567	0	0	49.9	18.8	27.3	315

### 4.1 Choice of Loss Function

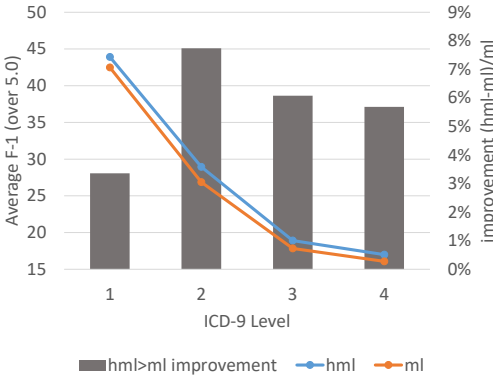
To assess the effect of using *hml* versus a standard multi-label loss function (*ml*) we examine all experimental results where the F1 was at least 5.0 (Fig. A.3). Models trained using *hml* consistently out-performed those trained using *ml* with an average F1 result between 3-8% better. This result holds when comparing the max values obtained in each levels as well with a 2-7% improvement for levels 2-4, although no significant improvement was seen for

## 4. Results

level 1. This last result is expected since the roll-up process for this level only rolls-up to level 0.

### 4.2 Discussion

The model's performance was able to recall 52.4% of the correct ICD-9 categories in a top-10 setting and assign the precise code in 82.6% of the cases. For this setting at the categorical level 0, and for the top-10 ICD codes (level 4), results are comparable to those published by Huang et al. [12] which predicted top-10 ICD-9 categories/codes by training deep neural networks over medical notes. In these days of automated electronic health records, this approach offers a potential application to automatically assign a disease code on the basis of drugs prescribed. This may also provide opportunities to create quality control mechanisms for diagnosis code assignment.



**Fig. A.3:** F1 difference between models trained using hierarchical multi-label loss (hml) and multi-label loss (ml).

the model was unable to find any of the cases ( $F1=0$ ). For example, at level 0, the model was unable to predict any assignment of chapters 780-799 (Symptoms, Signs, And Ill-Defined Conditions) and 710-739 (Diseases Of The Musculoskeletal System And Connective Tissue). These chapters may not be differentiable by medication, as the former is comprised of symptoms for many underlying conditions and the latter may be treated by orthopedic treatments and generic pain-relief medication. Further analysis shows that prediction of neoplasms mostly fails as well, as the treatment of cancer can be surgical or radiation-based. Furthermore, since MIMIC contains only ICU records, the patient may not be currently undergoing any medication-based cancer treatment. Further limitations include some drugs being prescribed for more than one diagnosis. For example, ACE inhibitors may be used for management of hypertension, heart failure, vascular disease and post-stroke. Also, doses of

F1 scores improve as the task is simplified with the worse performance obtained when the model is trying to assign the correct code from a set of 567 possible codes at level 4. The best performance is on level 0, when the model only has 15 possible labels. Consistently, in all experimental conditions, precision is higher than recall. This is partially explained by codes and groups that cannot be differentiated by their medication, and for which the

some drugs may vary depending on the disease indication, for example, rivaroxaban 2.5mg BID is licensed for high risk patients with acute coronary syndrome, while rivaroxaban 20mg OD is for stroke prevention in atrial fibrillation. Our analysis also does not consider changes in drugs over time, nor dose changes of the same drug. Also, some patients may swap their drug into another agent from the same class of drugs, causing a further dilution of the number of cases a model can learn from. Some drugs are also in combination therapies, for example, combining ACE inhibitors and a diuretic in a single *combo* pill for the treatment of hypertension.

## 5 Conclusion and Future Work

We presented a proof-of-concept study of the feasibility of using an ML-model to assign multiple diagnosis codes on multiple aggregation levels using a person’s current medication. The model was able to correctly assign diagnosis codes on multiple levels and the detailed results allow to identify which codes and code-groups are predictable by medication data. The use of a hierarchical loss function has improved the model’s performance by an average of 3–8%. The promising results support continued research into the ability to utilize larger medication datasets to create quality control mechanisms for diagnosis code assignment and to provide diagnostic information to caregivers in emergency situations that is language agnostic. We wish to pursue this expansion in future work, as well as experiment with additional hierarchical loss functions and methods to incorporate dosage and treatment regimen information in the model’s input.

## A Appendix - Omitted codes and detailed results

Table A.2 details the omitted codes from the diagnosis table and the reasons for omission. We omit all codes with a low number of cases. We further omit 61 codes used to describe symptoms, as these are shared by multiple causes and will, most-probably, supplant a diagnosis code following medical investigation. Injuries and foreign bodies (30 codes) are omitted as well as their treatment is usually orthopedic or surgical, rather than medicinal. We omit the codes used in ICD-9 to classify birth-age and pre-term phase for infants (14 codes) as these are more descriptive than diagnostic. Finally, we omit the E and V series of codes that are used to provide additional details for statistical reasons and which do not cause differences in medicinal treatment. We remain with 567 codes and 54,423 cases (92.4%) that contain at least one of the remaining codes. Filtering out only admissions contained in both the diagnosis and prescription tables we remain with 50,211 admissions.

## References

**Table A.2:** List of Omitted ICD-9 Codes and Code Groups

Code(s)	Description	Reason
5994 different codes	A large collection of various codes	Low base rate (less than 100 cases)
765.X	Descriptive of gestation week or preterm weight	Will be accompanied by the specific results of pre-term birth if such exist
8XX and 9XX	Injury	Medical result would be Surgical or Orthopedic and impossible to accurately specify from medication
93.31,93.41	Foreign body	Undiscernable medicinally
99.X	Complications of medical care	Undiscernable medicinally
61 different codes	Collection of different symptoms such as pain, nausea, and nuances of mental state/ faculties	Should be accompanied by the symptom's cause which is the main diagnosis

Detailed results are available online [10].

## References

- [1] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: Case study on ICD code assignment," in *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, ser. AAAI Workshops, vol. WS-18. AAAI Press, 2018, pp. 409–416. [Online]. Available: <https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16881>
- [2] R. Cerri, R. C. Barros, and A. C. De Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80 (1), no. 1, pp. 39–56, feb 2014.
- [3] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2018.
- [4] C. R. Cooke, M. J. Joo, S. M. Anderson, T. A. Lee, E. M. Udris, E. Johnson, and D. H. Au, "The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease," *BMC health services research*, vol. 11, no. 1, p. 37, 2011.
- [5] E.-M. Dalsgaard, D. R. Witte, M. Charles, M. E. Jørgensen, T. Lauritzen, and A. Sandbæk, "Validity of Danish register diagnoses of myocardial infarction

## References

- and stroke against experts in people with screen-detected diabetes." *BMC public health*, vol. 19, no. 1, p. 228, feb 2019.
- [6] G. Davie, J. Langley, A. Samaranayaka, and M. E. Wetherspoon, "Accuracy of injury coding under ICD-10-AM for New Zealand public hospital discharges," *Injury Prevention*, vol. 14, no. 5, pp. 319–323, oct 2008.
- [7] F. Fabris, A. A. Freitas, and J. M. Tullet, "An Extensive Empirical Comparison of Probabilistic Hierarchical Classifiers in Datasets of Ageing-Related Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 6, pp. 1045–1058, jan 2016.
- [8] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, sep 2016.
- [9] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [10] E. R. Hansen, T. Sagi, K. Hose, G. Y. H. Lip, T. B. Larsen, and F. Skjøth, "MIMIC Prescriptions result files," 2020. [Online]. Available: <https://doi.org/10.7910/DVN/5VTBME>
- [11] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics*, vol. 216, pp. 574–8, 2015.
- [12] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 141–153, aug 2019.
- [13] C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., sep 2017, pp. 3110–3113.
- [14] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, may 2016.
- [15] L. Kim, J.-A. Kim, and S. Kim, "A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples," *Epidemiology and Health*, p. e2014008, jul 2014.
- [16] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA*,



## References

- June 19-24, 2016, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1614–1623. [Online]. Available: <http://proceedings.mlr.press/v48/martins16.html>
- [17] A. Névél, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, “Clinical Natural Language Processing in languages other than English: Opportunities and challenges,” *Journal of Biomedical Semantics*, vol. 9, no. 1, mar 2018.
- [18] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, “Diagnosis code assignment: models and evaluation metrics,” *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 231–237, mar 2014.
- [19] N. Razavian, J. Marcus, and D. A. Sontag, “Multi-task prediction of disease onsets from longitudinal laboratory tests,” in *Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, ser. JMLR Workshop and Conference Proceedings, F. Doshi-Velez, J. Fackler, D. C. Kale, B. C. Wallace, and J. Wiens, Eds., vol. 56. JMLR.org, 2016, pp. 73–100. [Online]. Available: <http://proceedings.mlr.press/v56/Razavian16.html>
- [20] M. Schmidt, S. A. J. Schmidt, K. Adelborg, J. Sundbøll, K. Laugesen, V. Ehrenstein, and H. T. Sørensen, “The Danish health care system and epidemiological research: from health care contacts to database records,” *Clinical epidemiology*, vol. 11, pp. 563—591, 2019.
- [21] M. Schmidt, H. T. Sørensen, and L. Pedersen, “Diclofenac use and cardiovascular risks: Series of nationwide cohort studies,” *BMJ (Online)*, vol. 362, 2018.
- [22] S. A. Schmidt, M. Vestergaard, L. M. Baggesen, L. Pedersen, H. C. Schönheyder, and H. T. Sørensen, “Prevaccination epidemiology of herpes zoster in Denmark: Quantification of occurrence and risk factors,” *Vaccine*, vol. 35, no. 42, pp. 5589–5596, oct 2017.
- [23] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, “Clinical information extraction applications: A literature review,” *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, jan 2018.
- [24] J. Wehrmann, R. Cerri, and R. Barros, “Hierarchical Multi-Label Classification Networks,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 5075–5084. [Online]. Available: <http://proceedings.mlr.press/v80/wehrmann18a.html>
- [25] WHO, “International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10),” World Health Organization, Geneva, Switzerland, Tech. Rep., 2004.
- [26] R. Wockenfuss, T. Frese, K. Herrmann, M. Claussnitzer, and H. Sandholzer, “Three- and four-digit ICD-10 is not a reliable classification system in primary care,” *Scandinavian Journal of Primary Health Care*, vol. 27, no. 3, pp. 131–136, jan 2009.
- [27] D. Xu, Y. Shi, I. W. Tsang, Y. Ong, C. Gong, and X. Shen, “Survey on multi-output learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. (Early Access), pp. 1–21, 2019.

## References

- [28] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

# Paper B

## Assigning Diagnosis Codes using Medication History

Emil Riis Hansen, Tomer Sagi, Katja Hose, Gregory Y. H. Lip,  
Torben Bjerregaard Larsen, Flemming Skjøth

The paper has been published in the  
*Special Proceedings of the 18th International conference on Artificial Intelligence in  
Medicine (AIME 2020)*, vol 128, art. 102307, 2022.  
DOI: [10.1016/j.artmed.2022.102307](https://doi.org/10.1016/j.artmed.2022.102307)

## Abstract

*Diagnosis assignment is the process of assigning disease codes to patients. Automatic diagnosis assignment has the potential to validate code assignments, correct erroneous codes, and register completion. Previous methods build on text-based techniques utilizing medical notes but are inapplicable in the absence of these notes. We propose using patients' medication data to assign diagnosis codes. We present a proof-of-concept study using medical data from an American dataset (MIMIC-III) and Danish nationwide registers to train a machine-learning-based model that predicts an extensive collection of diagnosis codes for multiple levels of aggregation over a disease hierarchy. We further suggest a specialized loss function designed to utilize the innate hierarchical nature of the disease hierarchy. We evaluate the proposed method on a subset of 567 disease codes. Moreover, we investigate the technique's generalizability and transferability by (1) training and testing models on the same subsets of disease codes over the two medical datasets and (2) training models on the American dataset while evaluating them on the Danish dataset, respectively. Results demonstrate the proposed method can correctly assign diagnosis codes on multiple levels of aggregation from the disease hierarchy over the American dataset with recall 70.0% and precision 69.48% for top-10 assigned codes; thereby being comparable to text-based techniques. Furthermore, the specialised loss function performs consistently better than the non-hierarchical state-of-the-art version. Moreover, results suggest the proposed method is language and dataset-agnostic, with initial indications of transferability over subsets of disease codes.*

© The Authors 2022, published by Elsevier B.V under the CC-BY 4.0 License. Reprinted, with permission from Emil Riis Hansen, Tomer Sagi, Katja Hose, Gregory Y. H. Lip, Torben Bjerregaard Larsen and Flemming Skjøth.

Assigning Diagnosis Codes using Medication History. In: special proceedings of the 18th International conference on Artificial Intelligence in Medicine (AIME 2020), Volume 128, Article 102307, January 2022. <https://doi.org/10.1016/j.artmed.2022.102307>

*The layout has been revised.*

# 1 Introduction

The practice of coding diagnoses of medical conditions using standardized vocabularies of disease codes such as ICD-10 [2] has steadily grown. However, while coding systems are in widespread use, coding quality is uneven. Coding a medical diagnosis is notoriously complex. There exist multiple hierarchies and choosing the appropriate code requires a deep understanding of their structure and the relationships. For example, in a review of 1800 injury discharges from a New Zealand hospital, Davie et al. [7] found 2% to be uncoded, and 14% of principal injury diagnosis codes and 26% of external cause codes to be inaccurately coded. Wockenfuss et al. [33] determined that ICD-10 three and four level codes are too detailed to be reliable for general practitioners by measuring the Kappa inter-rater agreement scores.

Some work exists on predicting diagnoses from laboratory results (e.g., [23]), however, it is limited to cases where such results are available and relevant. A large body of work exists on extracting diagnoses from clinical notes and reports (see review [31]). However, the performance of these systems relies on techniques that tend to work much better in English and must be retrained for every new language [21].

A patient’s current medication can shed valuable light on their existing medical conditions. For example, observing that a patient has a long-term prescription for Metoprolol usually indicates that he/she is suffering from hypertension or ischaemic heart disease. Generalizing upon this observation, in this work, we develop a machine-learning-based model able to predict the list of diagnoses assigned to a patient based on his/her medications. Thus, such a model could provide emergency responders and critical care facilities with a rapid assessment of a patient’s existing conditions in addition to the model’s utility in diagnosis quality control. For example, an unconscious patient with a history of diabetes will be first assessed for hyper/hypoglycemia. In contrast, one without a history of diabetes but with a history of heart disease will be first assessed for acute heart conditions, such as a heart attack. We assess the viability of our approach using the publicly available American dataset (MIMIC-III) [16] and a Danish dataset combining prescription and diagnosis register data [18, 30] denoted DNPR in the following. While MIMIC-III contains rigorously anonymized and detailed medical records for over 50K intensive care unit (ICU) patients, DNPR contains data from an unselected population on disease codes from Danish hospital admissions and medication prescription history from Danish pharmacies.

This work extends our previous paper [26] in three ways. We investigate the generalizability and transferability of our approach by extensive experimentation on the Danish DNPR dataset. We investigate different aspects of heterogeneity between MIMIC-III and DNPR and provide results for compa-

rable non-medication-based methods.

The rest of the paper is structured as follows. In Section 2 we review related work. In Section 3 we describe MIMIC-III and DNPR, comparing and contrasting the two datasets. In Section 4 we detail our proposed method. In Section 5 we describe our experimental setup. In Section 6 we report experimental findings and provide results for text-based methods. We discuss the implications of the results in Section 7 while concluding and providing opportunities for future work in Section 8.

## 2 Related Work

Several studies justify the need to perform quality control of diagnosis code assignment. Cooke et al. [6] have shown that an ICD-9 code as a predictor of true chronic obstructive pulmonary disease had a sensitivity of 76% and specificity of 67% using spirometry as their gold standard. A comprehensive review of Danish validation studies on the Danish national patient registry [27] showed that the positive predictive values of disease and treatments varies from 15% to 100%. Recent work attempted to predict ICD-9 assignment in MIMIC-III from discharge notes [14]. Their solution to the multi-label multi-level problem was to limit the number of labels or aggregate predicted codes into categories, thereby solving two different problems, namely to predict the top-10/50 codes or the top 10/50 categories. In this work, we aim to predict a large set of codes at different aggregation levels to examine which codes and code groups are predictable from medication data.

There have been a few attempts to use prescription data to predict a single or at most two conditions. Schmidt et. al. developed and validated an algorithm with 87% accuracy able to identify herpes zoster [29]. In another study, prescription data was used to classify whether or not patients had preexisting conditions of diabetes or hypertension [28]. In a recent review [10] of algorithms designed to extract cases for medical research from electronic medical records data, some of the studies use medication data. However, all studies extract cases for a single condition, often aggregating several diagnosis codes. In our scenario, we identify the probable diagnosis codes of multiple conditions at once and thus identify cases where improbable diagnosis codes have been used.

## 3 Data and Heterogeneity

In this section we introduce the MIMIC-III and DNPR datasets and specify our steps of data preprocessing. Furthermore, to understand the heterogeneity between the datasets, we investigate and highlight their main differences.

#### 3.1 MIMIC-III

We use MIMIC-III [16] from PhysioNet [11], electronic health record (EHR) data for 50K patients who stayed in critical care units (ICU) of the Beth Israel Deaconess Medical Center for 11 years. MIMIC-III contains an extensive variety of data, including lab results, vital signs, medical notes, and most importantly for our needs, drugs administered, and diagnoses ascertained. MIMIC-III is structured as a relational database consisting of multiple tables. For instance, MIMIC-III contains a table for drug data, a table for diagnosis data, and a table for general patient information enclosing patient age, gender etc. The drug data table (model input) contains four million rows of drugs administered during 58,976 admissions. There are 4,525 different drug names in the DRUG field, which are often the same drug, with different spelling or with an added comment, e.g., *Basiliximab* and *\*NF\* Basiliximab*. To disambiguate and standardize the codes we use a mapping of MIMIC-III terms to the Observational Medical Outputs Partnerships (OMOP) Common Data Model (CDM) concepts [13] and group them by *Clinical Drug Form* to receive 1,602 RxNorm drug codes.

The diagnosis table (expected output) contains 651,047 diagnoses for 58,976 admissions using 6,984 different ICD-9 codes. ICD-9 is a hierarchical grouping of disease codes that consists of 5 levels starting from 0 (most general), to 4 (most specific). ICD-9 is built on the basis of grouping similar diseases. Upon review, we omit 6,110 codes for which less than 100 cases exist as it is typically not possible to generalize from such a low number. We further omit several codes focusing on diagnoses for persistent conditions not treatable by medication. A complete and detailed description of omissions can be found in A.

We use the patient table to add the *age* in years upon admission and *gender* to the model input normalized as described in Section 3.3.

#### 3.2 DNPR

To evaluate the generalizability of the proposed method, including its language-agnostic nature, we combine the two Danish datasets "The Danish National Patient Register" [18] and "The Danish National Prescription Registry" [30]. The Danish National Patient Register is the Danish national register of diagnosis data, which contains diagnosis codes assigned during patient hospitalizations. The register contains patient records since 1977. The Danish National Prescription Registry contains prescription data for all prescriptions sold in Denmark through pharmacies since 1994. The registers can be combined patient-wise through the Danish unique personal identification number which is used throughout in Danish registers. Demographic and vital status information is obtained from the Central Person Register [8]. Throughout this

work we refer to the combination of these three registers as DNPR. Due to the continuous approval of new drugs and expanding hierarchy of disease, we limit data from DNPR to the same range of years as MIMIC-III (2002 - 2012)

The combined register DNPR is structured as a relational database. It contains tables for patient diagnoses, prescribed drugs and general patient information among others. A main difference between MIMIC-III and DNPR is their different utilization of drug and disease vocabularies. Whereas MIMIC-III uses ICD-9 to code disease, DNPR uses a Danish extension of ICD-10 called The Danish Health Authority Classification System (SKS). Furthermore, DNPR utilizes the World Health Organization’s (WHO’s) Anatomical Therapeutic Classification (ATC) [25] for coding prescription drugs. DNPR contains 6,273,158 prescriptions (model input) and 2,351,769 diagnoses (expected output) for 2,093,987 admissions. Each admission (both inpatient and outpatient) consists of one or multiple diseases (both primary and secondary codes) diagnosed during hospitalization and all prescriptions administered to the patient within 30 days before and after diagnosis as illustrated in Figure B.1. In addition, we add patient *age* and *gender* to the model input normalized as described in Section 3.3.

### 3.3 Homogenization

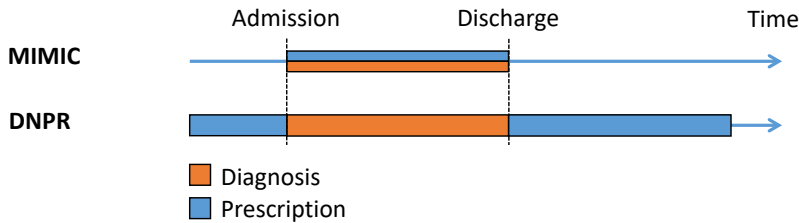
Due to the differences between MIMIC and DNPR, a homogenization of the datasets is required. As MIMIC-III and DNPR are coded using different disease and medication vocabularies, we created mappings to convert the DNPR disease and prescription codes to the code systems used in MIMIC-III. As detailed in Sections 3.4 and 3.4, based on many-to-many general equivalence mappings (GAMs) [3] and the OMOP CDM concept mappings, we managed to create a mapping for converting disease codes between ICD-9 and ICD-10-SKS as well as one for converting prescriptions from RxNorm to ATC. We map 567 unique ICD-9 codes to 320 unique ICD-10-SKS codes and 1602 unique RxNorm drug concepts to 834 unique ATC codes. Furthermore, MIMIC-III hides elderly patients (over 89 years) due to anonymization concerns and reports the age of 92.4 for each of these. We normalize the age of all patients from MIMIC-III by dividing it by 92.4; a practice that is beneficial in machine learning techniques. We normalize the age of patients from DNPR using the same approach by first calculating the average age of elderly patients (over 89 years), then reporting elderly patients with the average age, and finally dividing all patients by the average age of the elderly. When joining the prescription, diagnosis, and patient tables for MIMIC-III, we end up with 48K admissions for 38K different patients using 567 unique codes, referred to as labels in the following.



### 3.4 Data Heterogeneity

MIMIC-III and DNPR both consist of drug prescriptions and diagnosed diseases albeit they are collected for different purposes. Whereas MIMIC-III is collected in an insurance financed setting, DNPR is collected for administrative purposes but in a tax financed setting naturally leading to data heterogeneity.

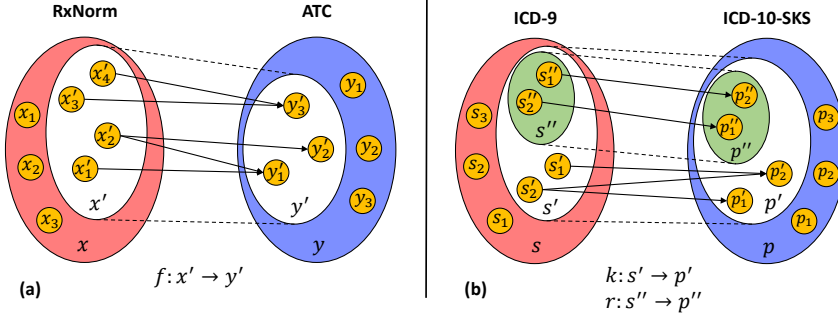
The main difference between MIMIC-III and DNPR is the way prescription data is gathered. While diagnosis codes from MIMIC-III and DNPR are both assigned while the patient is hospitalized, prescription data from DNPR differs from MIMIC-III by not consisting of the medicine administered during hospitalization but rather the medicine taken before and after release as illustrated in Figure B.1. Furthermore, since MIMIC-III consists of ICU patients often hospitalized with acute disease, the purpose of drug administration will initially be patient stabilization. On the other hand, the purpose of DNPR prescription data is directed at treating the disease diagnosed at release, as well as chronic conditions present before and after hospitalization (e.g., diabetes).



**Fig. B.1:** Differences between MIMIC-III and DNPR in terms of prescription data gathering. An orange box represents the time of diagnosis assignment and a blue box represents the time span for which prescription medicine consumption data is gathered. Whereas MIMIC-III contains information on prescription data from time of admission until release, DNPR only contains prescription data taken before and after the patient is released from the hospital.

#### Disease vocabularies

Although MIMIC-III and DNPR both utilize the ICD disease code hierarchy for standardized patient diagnosis, the hierarchy is used in different ways based on the purpose of the databases. Since subtle changes in disease codes can cause major changes to the final patient bill, MIMIC-III disease codes have to be as specific as possible. Comparatively, Danish physicians are not too concerned with the precision of specifying diagnosis codes as long as other clinicians can understand the patient's symptomatology. As an example, a patient from MIMIC-III might get diagnosed with the billable diagnosis code 280.1 - "Iron deficiency anemias - secondary to inadequate dietary iron



**Fig. B.2:** (a) RxNorm to ATC mapping. The mapping between RxNorm concepts and ATC codes forms a many-to-many relationship between the two vocabularies. This can be seen by the two RxNorm concepts  $x'_3$  and  $x'_4$  both mapping to the ATC concept  $y'_3$  and the RxNorm concept  $x'_2$  mapping to the two ATC concepts  $y'_1$  and  $y'_2$ . As an example, the RxNorm concepts "Digoxin Injection" and "Digoxin Oral Tablet" both map to the ATC concept "digoxin" and the ATC concepts "triamterene" and "hydrochlorothiazide" both map to the RxNorm concept "Hydrochlorothiazide / Triamterene Oral Tablet". (b) ICD-9 to ICD-10-SKS mapping. The mapping forms a many-to-many relationship between the two vocabularies as seen by subsets  $s'$  and  $p'$ . Furthermore, some codes only have one corresponding code from the other vocabulary, thus we create the sets  $s'' \subset s'$  and  $p'' \subset p'$  which have a one-to-one relationship. All mappings have been made available through an online data repository [12].

intake", whilst a patient from DNPR will be diagnosed with the less specific diagnosis code 280 - "Iron deficiency anemias", which is a non-billable ICD code.

For many years, ICD has been used globally and has thus gone through several iterations to accommodate new disease and better disease hierarchy structures. Whereas Denmark has been using the 10<sup>th</sup> version of ICD (ICD-10) since 1994, MIMIC-III patients have been diagnosed using the ICD-9 disease hierarchy. Furthermore, DNPR is coded using a Danish extension of ICD-10 called The Danish Health Authority Classification System (SKS) which extends the ICD-10 by introducing new branches of diseases and removing some codes that were originally in ICD-10. A bijective mapping between ICD-9 and ICD-10 is not possible due to the big changes between ICD versions [3]; however, a many-to-many mapping exists<sup>1</sup> as illustrated in Figure B.2(b) by the subset  $s'$  mapping to the subset  $p'$ . Additionally, we create subsets  $s'' \subset s'$  and  $p'' \subset p'$  of ICD-9 and ICD-10-SKS codes respectively for which there exists a one-to-one mapping between the sets; this is illustrated in Figure B.2 as the sets  $s''$  and  $p''$ . From the initial 567 ICD-9 codes with more than 100 MIMIC-III patient cases, we managed to map 320 unique ICD-9 codes to 532 ICD-10-SKS codes using the following procedure.

We utilize a many-to-many general equivalence mapping (GAM) between

<sup>1</sup><https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs>

### 3. Data and Heterogeneity

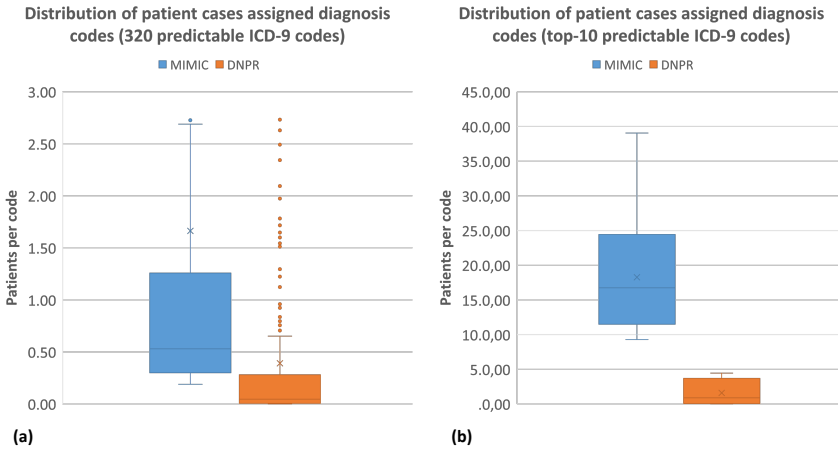
leaf nodes of the ICD-9 and ICD-10 disease hierarchies and consequently map 558 ICD-9 codes to 2525 ICD-10 codes. However, 1967 ICD-10 codes do not automatically correspond to ICD-10-SKS codes which results in 558 mappings from ICD-9 to ICD-10-SKS with 320 Unique ICD-9 codes mapping to 532 unique ICD-10-SKS codes forming a many-to-many relational mapping. Furthermore, we found a subset of 148 relations forming a one-to-one mapping between the two vocabularies.

#### Prescription vocabularies

Adding to the heterogeneous nature of prescription data, MIMIC-III and DNPR use different medicine vocabularies. Whereas MIMIC-III can be mapped to the RxNorm drug vocabulary using the OMOP CDM maps, DNPR is coded using the anatomical therapeutic classification (ATC). To compare the datasets we create a mapping from ATC codes to the RxNorm drug vocabulary using the OMOP CDM concept hierarchy. This results in a many-to-many mapping as seen in Figure B.2(a). Furthermore, the mapping is only partial since the OMOP CDM concept hierarchy has missing links between the two vocabularies. From the initial 1602 RxNorm drug codes, we were able to map 1257 unique RxNorm drug codes, illustrated as the set  $x' \subset x$  in Figure B.2, to 834 unique ATC drug codes, illustrated as the set  $y' \subset y$  in Figure B.2, with 1351 relations between  $x'$  and  $y'$ .

#### Statistical Heterogeneity

Of the resulting 834 mappable ATC codes, 771 are used at least once for patients from DNPR. Furthermore, counting only the 1,257 mappable drugs, the total number of drugs given to patients from MIMIC-III is 1,129,677 with an average of 23.72 drugs per patient case. In contrast, 6,273,158 drugs are prescribed to patients from DNPR averaging at 3.00 drugs per patient case. Likewise, using the many-to-many disease code mapping, we found that of the 320 unique ICD-9 disease codes, 307 have been assigned to patients from the DNPR dataset. MIMIC-III has 47,634 patient cases with a total of 282,150 assigned disease codes which gives an average of 5.94 diseases per patient. DNPR has 2,093,987 patient cases with a total of 2,351,769 diagnosed disease, averaging at 1.12 disease per patient. The distribution of patients diagnosed with each of the 320 mappable ICD-9 codes is illustrated in Figure B.3.



**Fig. B.3:** Distribution of patient cases with assigned ICD-9 codes. (a) The distribution of diagnosed patients for each of the 320 predictable ICD-9 codes. As illustrated, Q3 of DNPR codes is below the interquartile range of MIMIC-III codes. For the sake of readability, no percentage above 3 is shown. However, MIMIC-III has 29 outliers not shown on the figure and DNPR has 3. (b) The distribution of the top-10 used ICD-9 codes in MIMIC. As illustrated, all disease codes are used more frequently in MIMIC-III as compared to DNPR.

## 4 Hierarchical Multi-label Classification (HMC)

Binary classification problems (e.g., has this person received treatment related to sepsis) aim to correctly classify each task as either positive or negative. Single-label multi-class problems (e.g., is the following brain magnetic resonance imaging (MRI) normal or does it contain a glioblastoma, a sarcoma, or a metastatic bronchogenic carcinoma?) extend the classification to allow more than one class for each task. These two types of Machine Learning (ML) tasks are, by far, the most commonly studied in the medical domain. Less common are multi-label classification problems, which attempt to assign a set of labels to each example (e.g., which of the ICD-9 codes should be assigned following this medical report [1]), each of the labels is drawn from a possible set of classes. Since each person may have multiple co-morbidities, the task of assigning the correct set of diagnosis codes can be characterized as a multi-label classification problem [34]. The hierarchical nature of diagnoses both complicates the task and offers an opportunity to improve the applicability of an ML model. If an algorithm predicts a patient suffering from non-specified chiroisis (ICD-9 code 571.5) to be suffering from alcoholic chiroisis (ICD-9 code 571.2) it should be more appreciated than if no chiroisis related diagnoses are returned since both codes share a common ancestor. Further hierarchical constraints may dictate that a person cannot have more than one label from the same sub-tree of codes. Since ICD-9 is indeed hierar-

chical and imposes such constraints on some of its sub-trees, we can classify our task as a hierarchical multi-label classification (HMC) problem.

### 4.1 Machine Learning and Loss Functions

Many approaches to HMC include splitting the problem into multiple simple (single label) classification tasks, each of which is trained separately. Within these approaches, local and global approaches [9] differ by the number of classifiers trained. In the local case, multiple classifiers are trained over a binary label pertaining to a single node in the hierarchy and the predictions of each level are subsequently propagated [22]. In the global case, the labels are selected from a set of all possible labels. In this work, we follow the observation of Cerri et al. [4] that by training a single global classifier based on a multi-level neural network representation, one can effectively reuse the high-level features learned to discriminate between high levels in the hierarchy and then refine these to more accurate code assignments using the subsequent levels of the neural network. Furthermore, deep neural networks (DNN) have repeatedly shown superiority over other techniques in the medical domain (e.g., [15], [5]). We therefore employ a multi-layer perceptron, or fully connected neural network. The input layer for this network consists of one node for each RxNorm code in the data (one for normalized age and one for biological sex) and the output layer of one node for each ICD-9 code at the chosen roll-up level.

Machine learning, in particular deep learning, uses a loss function during the training phase to quantify the error of the current iteration of the model with respect to the expected output. Choosing an appropriate loss function is crucial and in general must reflect the structure of the expected output. Thus, specific loss functions have been suggested for the multi-label case [19] as well as hierarchical multi-label functions [32]. However, these are tied directly to the structure of the global classifier, and none have been applied in the medical data setting using the inherent hierarchy of a medical taxonomy.

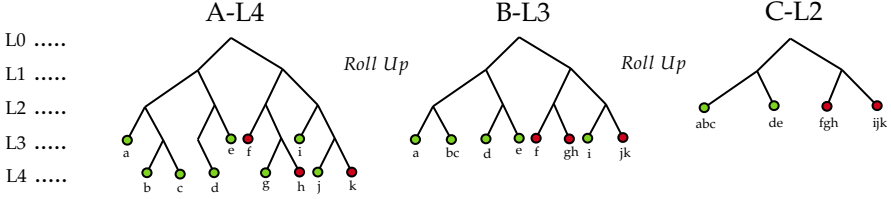
We therefore experiment with two types of loss functions, *ml* and *hml* as described below. One suitable for the multi-label case, where each missed label is treated the same regardless of the extent of the mistake (*ml*, Eq. 1), and one designed for the HMC case. For the general multi-label case, we chose the multi-label soft margin loss function [35], defined as follows with  $C$  being the number of classes,  $y$  being the class indicator, and  $x$  the current value of the corresponding output node ( $i$  iterates over all classes).

$$\begin{aligned} loss(x, y) = & -\frac{1}{C} \sum_i y[i] \cdot \log((1 + \exp(-x[i]))^{-1}) + \\ & (1 - y[i]) \cdot \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right) \end{aligned} \quad (\text{B.1})$$

We model our HMC loss function (*hml*, Eq. B.2) after the one developed for HMCN-F [32] while adjusting it to account for the differences between a text-classification problem and our own task and minimize a function comprised of two components.

$$\mathcal{L}_{hml} = \mathcal{L}_L + \mathcal{L}_G \quad (\text{B.2})$$

$\mathcal{L}_L$  is the local loss – calculation of Eq. B.1 at the leaf level.  $\mathcal{L}_G$  is calculated by rolling up the results one layer at a time until the ICD-9 chapter level (0). At each phase of the roll-up, the predictions for each inner node are set to the average of the predictions over its children. The loss of each level is calculated and summed to the other levels. Since our neural network does not directly predict the global scores, we do not suffer from hierarchical violations and do not require the third component that penalizes them in HMCN-F. We employ



**Fig. B.4:** Example of the roll up algorithm. An example level 4 code assignment is shown as tree A-L4. Disease codes {b, c, d, g, h, j, k} are level 4 codes, whereas codes {a, e, f, i} are codes on level 3. Red circles are the registered comorbidities of the patient. Green circles are diseases not recorded in the patient.

the Roll Up method to aggregate diagnoses given the ICD-9 hierarchy (see example in Figure B.4). Leaf node of the ICD-9 hierarchy can be assigned to patients. However, not all leaves are on the same level. As an example, 322.2 is a level 3 code, which represents *Chronic meningitis*, whereas code 003.22 is a level 4 code for *Salmonella pneumonia*. Each patient starts with one or more codes from the ICD-9 hierarchy.

## 5 Experimental Setup

In this section, we introduce the experimental setups for evaluating different aspects of our proposed method. We evaluate the proposed method’s overall performance by investigating the model’s performance on the MIMIC-III

## 5. Experimental Setup

dataset. Furthermore, we relate the model’s performance to baseline results from several textual-based diagnosis assignment methods. To evaluate the method’s generalizability, we investigate the model’s performance on the Danish DNPR dataset comparing it to the performance of MIMIC-III when trained and evaluated on the same sets of ICD-9 disease codes. Finally, we investigate the model’s transferability properties by training a model on the MIMIC-III dataset whilst testing the model on the DNPR dataset.

### 5.1 Diagnosis assignment using medication data (proposed method)

To evaluate the proposed method of using medication data to assign diagnosis codes, we train, evaluate and test *hml* and *ml* models on the MIMIC-III dataset with an 80/10/10 train/evaluate/test data split.

Utilizing the *roll up* method for initial data transformation, we perform experiments on different prediction resolutions, with level 0 corresponding to the chapter level of ICD-9 (e.g., 520–579: diseases of the digestive system) with 16 possible codes and level 1 to the code group level (e.g., 401-405 Hypertensive Disease) with 65 possible codes. Our last level corresponds to the most detailed available in the ICD-9 hierarchy (level 4) with 567 possible codes as identified in Section 3.1. Furthermore, we experiment with a Top-10 (level 4) setting and a Top-10 (Level 0) setting in terms of the most prevalent MIMIC-III codes. Furthermore, for each experiment, we perform a classic hyperparameter search over the number of internal layers and the number of nodes in each layer over the following values and ranges - *Activation function*: [Rectified Unit, Sigmoid], *Batch Size*: [32 - 2048], *layer Dropout*: [0.001 - 0.1], *Layer Sizes*: [1 - 4 layers, 128 - 512 perceptrons]. For each prediction resolution and parameter combination, we train and evaluate an *hml* and an *ml* model.

### 5.2 Generalizability

Generalizability should be understood as the model’s ability to perform well on new datasets and in new settings. To investigate the proposed method’s generalizability, we evaluate the model on the Danish DNPR dataset. DNPR is an ideal target for evaluating the method’s generalizability due to the heterogeneity between MIMIC and DNPR as detailed in Section 3.4. Since DNPR is coded using a different disease vocabulary and prescription vocabulary than that of MIMIC-III, the generalizability experiment reveals the dataset-agnostic and language-agnostic nature of the proposed method.

The experimental setup for the generalizability experiments are summarized in Table B.1. Experiments are performed using ICD-9 (level 4) codes on *hml* and *ml* models. Furthermore, all experiments are trained, evaluated,

Experiment	Codes	Disease Mapping	Train	Test
M - M (multi)	320	$k^{-1} : p' \rightarrow s'$	38,107	4,763
M - M (bijection)	148	$r^{-1} : p'' \rightarrow s''$	32,228	4,028
M - M (Top 50)	50	$k^{-1} : p' \rightarrow s'$	30,367	3,795
M - M (Top 10)	10	$k^{-1} : p' \rightarrow s'$	23,286	2,910
D - D (multi)	306	$k^{-1} : p' \rightarrow s'$	1,675,189	209,398
D - D (bijection)	142	$r^{-1} : p'' \rightarrow s''$	561,789	70,223
D - D (Top 50)	50	$k^{-1} : p' \rightarrow s'$	315,007	39,375
D - D (Top 10)	10	$k^{-1} : p' \rightarrow s'$	171,372	21,421

**Table B.1:** Experimental settings for evaluating the generalizability of the proposed method. M and D stand for MIMIC-III and DNPR respectively. Experiment is the name of the experiment where letters on the left and right side of the dash stand for the dataset used for training and testing respectively. Codes are the number of different disease predicted in the experiment. Train and Evaluation are the number of admissions for training and testing the model. Due to server limitations, hyperparameter optimization through standard grid search was not possible. Instead, model parameters were held constant for all experiments with the following settings - Batch Size: 256, Activation Function: Rectified Unit, Layer Dropout: 0.01, Layer Sizes: [512, 256, 128, 256]

and tested on an 80/10/10 data split. The DNPR data is hosted on a government server with severely restricted access and computational power thus limiting our ability to perform parameter grid search to tune the models. Hence, all experiments use the same parameter settings which can be found in the legend of Table B.1.

### 5.3 Transferability

Transferability is the model’s ability to work in a different setting from the setting in which it has been originally trained. We evaluate the proposed method’s transferability by training a model on the MIMIC-III dataset while testing the model on the DNPR dataset. The transferability experiments are listed in Table B.2. To ensure data compatibility, we preprocess the DNPR dataset by translating the model input and output according to the taxonomy mappings developed in Section 3.4 and Section 3.4 as illustrated in Figure B.2. All experiments are done for the most detailed level of ICD-9 (level 4) using both an *hml* and *ml* model. Due to server limitations, model parameter optimization is not possible. Model parameter settings are held constant, as listed in Table B.2. All transferability experiments are trained and evaluated on an 80/20 MIMIC data split while tested on all DNPR data.



## 5.4 Experimentation Settings

For the generalizability and transferability experiments as described in Sections 5.2 and 5.3, we experiment with multiple settings of disease codes and hierarchies. Each setting has a different rationale and clinical application in hospital settings.

The multi experiments utilize the  $k^{-1}$  disease mapping as described in Section 3.4.  $k^{-1}$  establishes a many-to-many link between diseases of the ICD-9 vocabulary and that of the ICD-10 vocabulary. In total, we were able to map 320 ICD-9 codes to ICD-10 codes using this mapping. The mapping is a naive conversion method since the mapping from ICD-10 to ICD-9 merges several ICD-10 codes into a single ICD-9 code. However, since most groups of merged ICD-10 codes are very similar, it should be uncommon for patients to lose important disease information when using the mapping. Furthermore, this mapping keeps many of the original disease codes from the 567 ICD-9 code set. A model based on such a mapping can be used in a clinical setting for various purposes such as automatic disease code assignment, as a validation tool for manual disease code assignment, for finding registry errors, or as a clinical tool for assessing the disease history of a patient based on the patients prescription history.

To evaluate the performance of one-to-one corresponding codes from ICD-9 and ICD-10, we created a bijective mapping function  $R^{-1}$  to map ICD-10 disease codes to ICD-9 disease codes. The experiments using this mapping are mainly used to investigate the performance of a model when mitigating the problems introduced by many-to-many mappings.

Top 10 (level 4) and top 50 (level 4) experiments use the top 10 and top 50 diagnosed codes. Previous diagnosis assignment approaches [17, 24] have used top 10 and top 50 codes for experimentation. To be comparable with other approaches for diagnosis assignment on the MIMIC-III dataset, we chose to incorporate these experimental settings as well.

Experiment	Codes	Disease Mapping	Train	Test
M - D (multi)	320	$k^{-1} : p' \rightarrow s'$	47,634	2,093,987
M - D (bijection)	148	$r^{-1} : p'' \rightarrow s''$	40,286	693,950
M - D (Top 50)	50	$k^{-1} : p' \rightarrow s'$	37,959	389,344
M - D (Top 10)	10	$k^{-1} : p' \rightarrow s'$	29,108	211,579

**Table B.2:** Experimental settings for evaluating the transferability of the proposed method. M and D stand for MIMIC-III and DNPR, respectively. The description of the legend and experiments follows the same format as that of Table B.1.

## 5.5 Comparison to text-based methods

We evaluated several textual-based approaches similar to those proposed by [24] for diagnosis assignment on different sets of the 567 MIMIC-III codes described in Section 3.1. We evaluated a Convolutional Neural Network (CNN) [20], a Recurrent Neural Network followed by a Gated Recurrent Unit (GRU), and a Convolutional Neural Network with Attention (CNN-att) [24].

The evaluated text-based methods treat ICD-9 code prediction as a multi-label classification problem. The input for text-based methods are the textual discharge summaries for patient stays, and the output is the ICD-9 codes assigned to the patient. To compare against our approach, we evaluate each of the three text-based models in a Top-10 (level 4) setting and a Raw (Level 4) setting, with Top-10 occurring MIMIC-III (level 4) codes and the set of all 567 MIMIC-III (level 4) codes, respectively.

The convolutional neural network we evaluate against, as described in [24], works as follows. As an initial data transformation step, the discharge summary notes are transformed into a feature matrix by substituting each word using pre-trained  $d_e$ -dimensional word embeddings to create an embedding matrix  $X = [x_1, x_2, \dots, x_N]$ , where  $N$  is the length of the document. A convolution layer then applies a convolutional filter  $W_c \in \mathbb{R}^{k \times d_e \times d_c}$ , where  $d_c$  is the size of the filter output, to  $X$ , to produce a convolution matrix  $H$ . A global average pooling layer is then applied to  $H$  to generate a feature for each corresponding disease to classify. The only difference between the CNN and the GRU network architecture is that a gated recurrent unit layer replaces the convolution layer from the CNN-based architecture. The CNN-att model utilizes a per-label attention mechanism since different parts of the convolution  $H$  may be relevant for different labels. The attention mechanism learns a vector parameter  $u_l \in \mathbb{R}^{d_c}$  for each disease label. By doing matrix multiplication between  $u_l$  and  $H$  and using a *softmax* function to normalize over all words from the input file, an attention vector  $a_l$  is learned for each label. The intuition behind  $a_l$  is that it learns which words in a document are important for classifying a specific label  $l$ .

## 5.6 Baseline

We introduced a statistics based disease code assignment approach as a baseline method for the task of disease code assignment. The approach is based on the statistical prior that patients are more likely to be diagnosed with common diseases than rare diseases. For each disease, we first calculate the dataset-specific probability of a patient having a disease. The assignment of patient diseases then follows a schema of generating a random floating point number between 0 and 100 for each patient for each disease. If the randomly generated number is lower than or equal to the probability of having the

disease, we assign the diagnosis code to the patient. A good model should outperform this baseline by learning from the input features to choose against the statistical prior.

## 6 Experimental Results

This section presents the results obtained from the experimental settings defined in Section 5. The obtained results are presented in separate sections according to their experimental setting. To allow easy comparison between our approach and techniques utilizing medical notes, we evaluate experimental results using the standard micro-averaged precision and recall and their harmonic mean F1. The choice of experimental settings is described in Section 5.4.

### 6.1 Diagnosis assignment using medication data (Proposed Method)

To evaluate the proposed method, we trained several models on the MIMIC-III dataset for each ICD-9 level according to the experimental setup described in Section 5.1. Table B.3 presents the best results (by F1) obtained over MIMIC-III using an 80/10/10 split by an

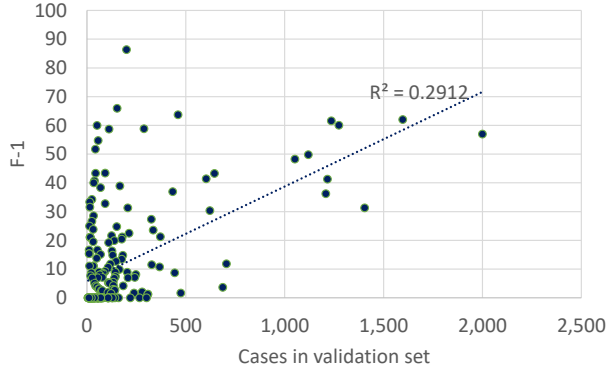


Fig. B.5: F-1 by number of cases over level 2 codes.

*hml* mode following a standard hyper-parameter grid search. In each task, the code assignments were rolled up before both the training and the test phase and not only for evaluation, such that the neural network encountered a different task for each level. For each ICD-9 level, we provide the number of codes in that level, the average branching factor, and the average number of eventual leaves of a node in this level's sub-tree. In addition to precision, recall, and F1, we show the number of diagnosis codes for which F1 was equal to zero. Table B.3 further presents the results of the baseline approach for easily comparing our proposed method against the baseline. We evaluate the Raw (level 4), Top-10 (level 0) and Top-10 (level 4) tasks for the baseline.

Prediction Task	Codes	Br.	Avg. Leaves	Prec.	Recall	F1	F1=0
<b>Baseline</b>							
Top-10 (level 0)	10	NA	NA	42.04	40.10	41.05	1
Top-10 (level 4)	10	NA	NA	23.25	21.46	22.32	0
Raw (level 4)	567	0	0	8.79	8.24	8.51	402
<b>Our Approach</b>							
Top-10 (level 0)	10	NA	NA	69.48	70.23	70.01	0
Top-10 (level 4)	10	NA	NA	52.38	70.00	59.92	0
Rolled Up (level 0)	16	5.7	565.1	68.46	69.27	68.86	0
Rolled Up (level 1)	65	8.4	108.3	58.05	57.21	57.63	10
Rolled Up (level 2)	236	6.6	14.0	48.45	47.19	47.81	83
Rolled Up (level 3)	461	1.6	1.6	37.36	41.61	39.37	195
Raw (level 4)	567	0	0	36.98	36.26	36.62	311

**Table B.3:** MIMIC-III diagnosis prediction results for our approach and for the baseline. Br. is the branching factor and Prec. is precision. F1=0 is the number of codes for which F1 was equal to zero.

Since MIMIC-III is a relatively small dataset, the number of cases for many diagnoses is too low to expect good performance. When examining the effect of the number of cases on the model’s performance (Fig. B.5) we find that at least some of the variance can be explained by the small number of cases ( $R^2$  of 0.29 for a linear model). Top-5/top-10 results by code are available as an online appendix containing the full results [12].

To assess the effect of using a hierarchical multi-label loss function (*hml*) versus a standard multi-label loss function (*ml*) we examine all experimental results from the *proposed method* experiment as described in Section 5.1 where the F1 was at least 5.0. Models trained using *hml* consistently out-performed those trained using *ml* with an average F1 result between 3 – 8% better. This result holds when comparing the max values obtained in each level with a 2 – 7% improvement for levels 2 – 4, although no significant improvement was seen for level 1. This last result is expected since the roll-up process for this level only rolls up to level 0.

## 6.2 Generalizability Results

To investigate the proposed method’s generalizability, we compare the performance of models trained on the MIMIC-III dataset to models trained on the same set of ICD-9 codes on the Danish DNPR dataset. The experimental setting is described in section 5.2. Results in terms of F1 scores for *hml* and *ml* models grouped by experimental setting for all generalizability experiments

## 6. Experimental Results

are illustrated in Figure B.6.

Even though the two datasets are heterogeneous in nature, as described in Section 3.4, results indicate that the proposed method provides comparable predictive power for models trained on the MIMIC-III dataset and models trained on the DNPR dataset, for the same subsets of ICD-9 codes. Inter-

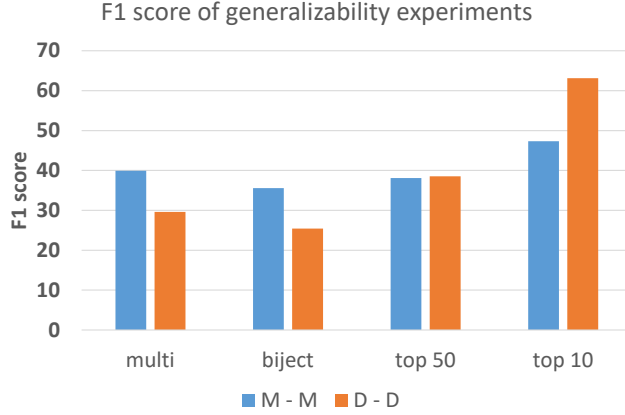


Fig. B.6: F1 scores for *hml* models grouped by the type of experiment.

estingly, models trained on MIMIC-III outperform models trained on DNPR when the number of predictable diseases is high. In contrast, the opposite is true when the number of predictable diseases is low.

Furthermore, results obtained from the generalizability experiments further validate the superiority of using an *hml* model as illustrated by Figure B.7, as *hml* models persistently out-performed *ml* models on F1 scores by up to 7.5% with an average performance increase of 3.1%.

Experiment	Codes	HML			ML		
		F1	Prec.	Recall	F1	Prec.	Recall
M - M (multi)	320	39.93	40.16	39.71	37.30	36.53	38.12
M - M (bijection)	148	35.56	34.64	36.53	31.34	29.46	33.48
M - M (Top 50)	50	38.09	36.51	39.81	34.83	32.72	37.23
M - M (Top 10)	10	48.62	40.93	59.86	41.16	40.83	41.49
D - D (multi)	306	30.25	31.70	28.92	30.24	30.64	29.86
D - D (bijection)	142	25.42	20.43	33.61	24.42	24.36	26.57
D - D (Top 50)	50	38.55	36.25	41.16	38.09	35.14	41.58
D - D (Top 10)	10	63.16	59.54	67.25	57.38	52.42	63.38

Table B.4: F1, precision and recall of generalizability experiments for *hml* and *ml* models. M and D stand for MIMIC-III and DNPR respectively.

### 6.3 Transferability Results

To assess the proposed method’s transferability, we performed experiments described in Section 5.3. We trained and evaluated a model on the MIMIC-III dataset for each transferability experiment with an 80/20 data split while testing the model on the whole DNPR dataset. Results in terms of F1 score, precision and recall for all transferability experiments

for *hml* and *ml* models are presented in Table B.2. Results in terms of F1 score range from 6.28 when trained and tested on 320 disease codes to 28.25 when trained and tested on the top-10 most prevalent MIMIC-III ICD-9 codes as summarized in Table B.5. Although transferability results indicate weak performance for models trained on the 320 ICD-9 codes, the performance improves as the prediction task gets easier.

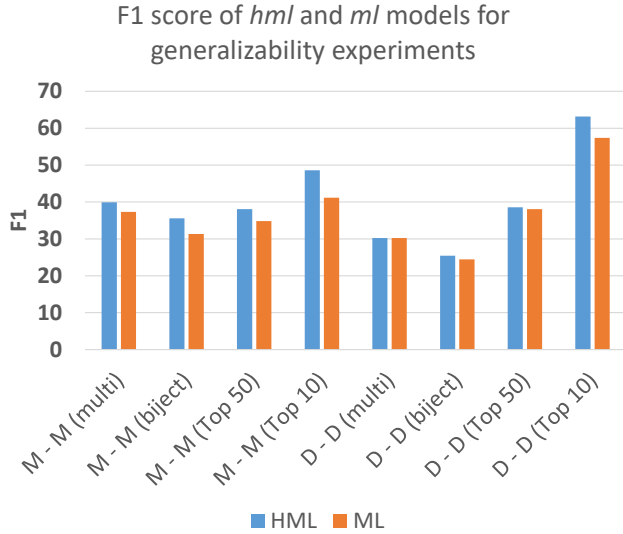


Fig. B.7: F1 scores for generalizability experiments as listed in Table B.1

Experiment	Codes	HML			ML		
		F1	Prec.	Recall	F1	Prec.	Recall
M - D (multi)	320	6.28	7.38	5.46	5.46	4.72	6.49
M - D (bijection)	148	6.68	4.77	11.12	5.92	3.75	13.90
M - D (Top 50)	50	10.86	7.65	18.70	9.70	6.27	21.29
M - D (Top 10)	10	28.25	19.26	50.45	21.22	20.34	22.17

Table B.5: F1, precision and recall of transferability experiments for *hml* and *ml* models. M and D stands for MIMIC-III and DNPR respectively.

## 6.4 Results for text-based methods

To compare our work to text-based methods of diagnosis assignment, we experimented with implementations of several such methods. We evaluated state of the art Convolutional Neural Network (CNN), a Recurrent Neural Network followed by a Gated Recurrent Unit (GRU), and a Convolutional Neural Network with Attention (CNN-att) [24]. Results in terms of precision,

## 7. Discussion

recall, and F1 for all textual techniques are listed in Table B.6. The best result in terms of F1 for the Top 10 (level 4) experiment was achieved with a CNN-att model with a score of 82.74. In comparison, the best result for our proposed method on the same set of codes is 59.92 as listed in Table B.3. Similarly, whereas our proposed method achieved an F1 score of 36.62 when predicting the complete set of 567 ICD-9 codes (level 4), the best result for the text-based methods was achieved on the CNN-att model with a score of 55.76.

Model	Codes	Precision	Recall	F1
CNN	Top 10 (level 4)	76.13	77.65	76.89
CNN	Raw (level 4)	44.51	46.33	42.82
GRU	Top 10 (level 4)	77.82	82.65	80.16
GRU	Raw (level 4)	62.02	49.96	55.34
CNN-att	Top 10 (level 4)	79.34	82.26	80.77
CNN-att	Raw (level 4)	57.51	59.38	55.76
<i>hml</i>	Top 10 (level 4)	<b>54.24</b>	<b>67.92</b>	<b>60.31</b>
<i>hml</i>	Raw (level 4)	<b>36.98</b>	<b>36.26</b>	<b>36.62</b>

**Table B.6:** Results of text-based methods of diagnosis assignment. Codes are the sets of ICD-9 disease codes used in the experiment. Top 10 (level 4) is the 10 most frequently used ICD-9 codes in MIMIC-III from the initial set of 567 codes. Raw (level 4) is the complete set of 567 ICD-9 codes. For comparison, the table contains the best results for medication-based diagnosis code assignment for the same tasks.

We further present the results for the Top-10 (level 4) assigned diagnosis codes for the text-based CNN model and our medication based HML model in Table B.7. CNN predicts "Atrial fibrillation" with an F1 score of 89.66 whereas HML predicts the same disease with an F1 score of 73.09. The best performing class in terms of F1 score for the HML model is "Coronary atherosclerosis of native coronary artery" with an F1 score of 68.84. CNN predicts the same code with an F1 score of 77.23. Further results with the Top-10 assigned codes for the GRU and CNN-att text-based models are available as an online appendix [12].

## 7 Discussion

This section discusses and reflects upon the experimental results for the proposed method and its generalizability and transferability.

Disease	HML			CNN		
	Prec.	Recall	F1	Prec.	Recall	F1
Atrial fibrillation	69.03	77.65	73.09	87.62	91.78	89.66
Coronary atherosclerosis of native coronary artery	63.05	75.79	68.84	92.81	66.13	77.23
Unspecified essential hypertension	55.54	85.65	67.38	70.72	90.84	79.53
Congestive heart failure; unspecified	60.20	72.92	65.95	86.10	79.20	82.50
Acute respiratory failure	51.77	73.02	60.58	66.27	66.93	66.60
Acute kidney failure; unspecified	45.23	62.50	52.48	77.48	45.43	57.27
Diabetes mellitus without mention of complication	49.09	56.28	52.44	71.28	82.75	76.59
Urinary tract infection; site not specified	42.09	61.58	50.00	71.68	70.13	70.90
Other and unspecified hyperlipidemia	44.72	49.48	46.98	77.96	76.26	77.10
Esophageal reflux	36.77	21.32	26.99	82.10	67.19	73.90

**Table B.7:** F1, precision and recall of top-10 assigned ICD-9 codes for our medication-based *hml* model and the text-based CNN model.

## 7.1 Proposed method

In the top-10 setting, an *hml* model was trained to assign one or more diseases to a patient among 10 unique ICD-9 disease codes. The model correctly assigned codes in 69.48% of all cases and was able to find 70.23% of all disease codes as summarized in Table B.3.

The results of the performance of the top-10 assigned codes setting shows that text-based methods perform well in the diagnosis of all top-10 diseases as summarized in Table B.7. The results indicate that diagnosis observations are diligently written down in clinical discharge notes, with a precision such that text-based methods of diagnosis classification works well. Not surprisingly, it is more difficult to differentiate between diagnosis codes based on medication since some medications can be used in various contexts for treating multiple diseases. Furthermore, some diseases are not treated directly, but by adjusting some other treatments if the disease is a side effect, such as is often the case with Esophageal reflux. Hence, the F1 score of 26.99 for the medication based prediction of Esophageal reflux. However, for 8 out of 9 top-10 assigned codes, the F1 score for our medication based HML model was above 50.

The results are encouraging compared to the CNN, GRU and CNN-att



## 7. Discussion

textual methods of diagnosis assignment as illustrated in Figure B.6. In these days of computerized electronic health records, this approach offers a potential application to assign disease codes based on drugs prescribed automatically. The approach may also provide opportunities to create quality control mechanisms for diagnosis code assignments. The proposed method works in cases where registers do not contain medical notes but contain patient medication history, as in the Danish patient register DNPR.

As summarized in Table B.3, F1 scores improve as the task is simplified with the worse performance obtained when the model tries to assign the correct code from a set of 567 possible codes at level 4. The best performance is on level 0 when the model only has 16 possible labels. Consistently, in all experimental conditions, precision and recall are approximately the same. Precision and recall are relatively low when predicting all 567 (level 4) codes. This result is partially explained by codes and groups that their medication cannot differentiate, and for which the model was unable to find any of the cases ( $F1=0$ ). For example, at level 4, the model could not predict any assignment of codes from chapter 780-799 (Symptoms, Signs, And Ill-Defined Conditions). This chapter may not be differentiable by medication, as it comprises symptoms for many underlying conditions. Further analysis shows that prediction of neoplasms mostly fails, as cancer treatment can be surgical or radiation-based. Furthermore, since MIMIC contains only ICU records, the patient may not be currently undergoing any medication-based cancer treatment.

In addition, many diseases of the circulatory system were not differentiable by medication. Some diseases are asymptomatic and will thus rarely be treated by medication since the patient does not produce or show any symptoms regardless of the presence of the disease. The branch of diseases under code 426 (Conduction Disorders) are mostly asymptomatic, such as 426.0 (Atrioventricular Block, Complete), 426.4 (Right Bundle Branch Block), and 426.7 (Anomalous Atrioventricular Excitation). Other diseases are either too general, as in 427.89 (Other Specified Cardiac Dysrhythmias), which makes it medically undiscernible, or does not have a specific medication treatment regime such as 437.0 (Cerebral Atherosclerosis). The treatment of cerebral atherosclerosis often involves administering statin, used for lowering cholesterol levels in the blood. However, statin is also used for various other atherosclerosis diseases such as aortic atherosclerosis and atherosclerosis of renal artery. Since no other discernable medication is used to treat cerebral atherosclerosis, this disease cannot be differentiated by medication.

Nonetheless, in some cases, diseases will have specific regimes of medication treatment, such as atrial fibrillation and hypertension. Patients with hypertension will often be treated by beta-blockers, ACE inhibitors or angiotensin II inhibitors. If two of these have been prescribed to a patient, there is a high probability of suffering from hypertension.

Another issue that is difficult to capture is that doses information of some drugs may vary depending on the disease indication. For example, rivaroxaban 2.5mg BID is licensed for high-risk patients with acute coronary syndrome, while rivaroxaban 20mg OD is for stroke prevention in atrial fibrillation. In this paper, we focused our analysis on static patient information, which means that we do not model changes in drugs over time. For example, the medication warfarin will often be prescribed to patients with venous thromboembolism and patients with atrial flutter. Whereas patients with atrial flutter will be prescribed warfarin for their entire life, venous thromboembolism patients will often stop taking warfarin after a certain period. Designing a model that can capture temporal drug information is an interesting aspect that we plan to address in our future work. Also, some patients may swap their drug into another agent from the same class of drugs, causing a further dilution of the number of cases a model can learn from. Some drugs are also in combination therapies, for example, combining ACE inhibitors and a diuretic in a single *combo* pill for the treatment of hypertension.

## 7.2 Generalizability

We evaluated the generalizability of the proposed method by experimenting with the Danish DNPR dataset. We compared results obtained from *hml* and *ml* models created over sets of ICD-9 codes from the MIMIC-III dataset to results obtained over the same sets of ICD-9 codes from the DNPR dataset. Experiments are summarized in Table B.1. Despite their different aspects of heterogeneity, experimental results indicate comparable predictive model power for both datasets as illustrated in Figure B.6. This finding demonstrates the proposed method’s dataset-agnostic properties. Furthermore, as the Danish and American datasets use distinct prescription and diagnosis vocabularies with different naming conventions for medications and diseases, we created mappings to convert between the vocabularies as described in Sections 3.4 and 3.4. Even though the mappings are incomplete and include many-to-many relations, results indicate that such conversion does not hurt the predictable properties of the proposed model when used on the Danish dataset. This result demonstrates the proposed method’s language-agnostic properties.

## 7.3 Transferability

As indicated by the results gained from investigating the model’s transferability, patient data’s heterogeneous nature negatively affects the proposed methods predictive power. We evaluated the transferability of the proposed method by training models on subsets of ICD-9 disease codes of the MIMIC-III dataset while evaluating the models on the same sets of ICD-9 codes for

the DNPR dataset as listed in Table B.2. Results are summarized in Table B.5. We achieve the F1 score of 6.28% from training an *hml* model on 320 ICD-9 level 4 codes while testing on the same subset of ATC-converted ICD-9 codes from the DNPR dataset. Furthermore, for 229 out of 320 disease codes, the model could not provide any accurate predictions ( $F1=0$ ). The results suggest that the heterogeneity between patient data across countries is too considerable to create a model with good transferability. As investigated in Section 3.4, the variability in purpose, collection method, and utilization of diverse vocabulary standards for prescription and disease code hierarchies arguably add to the variance between MIMIC and DNPR. Notwithstanding, when limiting to subsets of ICD-9 codes, model transferability significantly improves. An *hml* model trained on the top 10 occurring ICD-9 codes from the  $s'$  code subset achieves an F1 score of 28.25 when tested on the same 10 ATC converted codes from the DNPR dataset. Noticeably, 4 out of 10 ICD-9 codes achieve an F1 score below 5.00, which indicates that the proposed method could potentially have a high transferability on specific sets of disease codes.

## 7.4 Domain Knowledge

As with the majority of AI models today, domain knowledge is required to train models with satisfactory performance in real world applications. Although the proposed method incorporates external knowledge such as the ICD-9 disease code hierarchy and the RxNorm medication vocabulary, the proposed method is in fact agnostic towards these. Given a medical dataset coded using arbitrary disease and medication vocabularies, one could train a model using the proposed method either with or without a hierarchical taxonomy over the vocabularies. While our model performs adequately without any added domain knowledge, we show that incorporating domain knowledge in the form of hierarchical taxonomies directly into the loss function for multi-label diagnosis prediction consistently improves model results.

## 7.5 Practical Implications

Automatic diagnosis code assignment using medication history has multiple practical implications such as registry error correction, a supportive validation tool for manual code assignment, or indicative tools usable in cases where prescription information is present but diagnosis information is not. Disease registers with manually assigned disease codes have been shown to be error-prone [7]. Using a neural model to find general patterns of medication to disease indications could automatically find outliers in register data. Currently we achieve an F1 score of 36.98% on a model for the prediction of 567 codes as summarized in Table B.3. Furthermore, 311 of these codes are

not discernible by medication. Hence, a neural model using only patients' prescription history can not find registry errors for all disease codes. However, as our experiments show, medication history can for some diagnosis codes be used for highly accurate diagnosis prediction and thereby be used in a system for finding register errors.

Manual diagnosis assignment is a cumbersome and error-prone task. Using a supportive tool to validate medical inputs of clinicians could help catch errors before they enter the system. As summarized in Table B.3 the model performs better on higher levels of prediction. Although the model might not catch wrongly assigned diagnosis codes at the most specific level (level 4), it could help catch cases on higher aggregation levels where the error's severity is large.

In countries such as Germany, disease and medication registers are not combined. This means that emergency health care providers in ambulant settings sometimes only know the patient's prescription history and not the disease history. This can have severe implications for the treatment of the patients, such as in the case of a patient having diabetes where several treatment protocols drastically change. In this case, a medication based diagnosis prescription model could help identify serious diseases present in the patient to guide emergency health care providers in providing the correct treatment protocol in ambulant settings.

## 8 Conclusion and Future Work

We presented a proof-of-concept study of the feasibility of using a machine learning model to assign multiple diagnosis codes on multiple aggregation levels using a person's current medication. The proposed method correctly assigned diagnosis codes on multiple levels of the ICD-9 hierarchy over the MIMIC-III dataset. The detailed results allow identifying which codes and code-groups are predictable by medication data. The use of a hierarchical loss function improved the proposed method's performance by an average F1 of 3-8% on multiple levels of aggregation of the MIMIC-III dataset while also increasing generalizability results by up to 7.5% in terms of F1 score. The promising results support continued research into utilising larger medication datasets to create quality control mechanisms for diagnosis code assignment and provide diagnostic information to caregivers in emergencies.

Future work will further explore applications to clinical care using medication based diagnosis. Generalizability experiments demonstrate the feasibility and efficiency of the technique when applied to new dataset. Generalizability results from experimentation on the Danish DNPR dataset indicate that the technique is language-agnostic and can be directly used over new datasets. The technique is also helpful in situations where prescription data

is present, but clinical discharge notes are not, as is the case with DNPR. Although model transferability underperformed when tested on the Danish DNPR dataset, results indicate that specific subsets of codes could be trained to perform well, even in model transferability. Furthermore, integrating more and diverse patient information into a unified model for diagnosis prediction should be further investigated. Patient clinical notes, medical imaging, coding systems such as laboratory codes, symptom codes and others are but a few examples of the diverse information contained in patient EHR that combined could increase the predictive performance of medical AI systems.

## A Appendix - Omitted codes and detailed results

Table B.8 details the omitted codes from the diagnosis table and the reasons for omission. We omit all codes with a low number of cases. We further omit 61 codes used to describe symptoms, as these are shared by multiple causes and will, most-probably, supplant a diagnosis code following medical investigation. Injuries and foreign bodies (30 codes) are omitted as well as their treatment is usually orthopedic or surgical, rather than medicinal. We omit the codes used in ICD-9 to classify birth-age and pre-term phase for infants (14 codes) as these are more descriptive than diagnostic. Finally, we omit the E and V series of codes that are used to provide additional details for statistical reasons and which do not cause differences in medicinal treatment. We remain with 567 codes and 54,419 cases (92.4%) that contain at least one of the remaining codes. Filtering out only admissions contained in both the diagnosis and prescription tables we remain with 48,516 admissions.

## References

- [1] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: Case study on ICD code assignment," in *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, ser. AAAI Workshops, vol. WS-18. AAAI Press, 2018, pp. 409–416.
- [2] G. Brämer, "International statistical classification of diseases and related health problems. tenth revision," *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, vol. 41, no. 1, p. 32–36, 1988.
- [3] D. J. Cartwright, "Icd-9-cm to icd-10-cm codes: what? why? how?" in *Advances in wound care*. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, 2013, pp. 588–592.
- [4] R. Cerri, R. C. Barros, and A. C. De Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80 (1), no. 1, pp. 39–56, feb 2014.

## References

**Table B.8:** List of Omitted ICD-9 Codes and Code Groups

Code(s)	Description	Reason
5994 different codes 765.X	A large collection of various codes Descriptive of gestation week or preterm weight	Low base rate (less than 100 cases) Will be accompanied by the specific results of pre-term birth if such exist
8XX and 9XX	Injury	Treatment would be Surgical or Orthopedic and impossible to accurately specify from medication
93.31,93.41 99.X	Foreign body Complications of medical care	Undiscernable medicinally Undiscernable medicinally
61 different codes	Collection of different symptoms such as pain, nausea, and nuances of mental state/faculties	Should be accompanied by the symptom's cause which is the main diagnosis

- [5] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2018.
- [6] C. R. Cooke, M. J. Joo, S. M. Anderson, T. A. Lee, E. M. Udris, E. Johnson, and D. H. Au, "The validity of using icd-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease," *BMC health services research*, vol. 11, no. 1, pp. 1–10, 2011.
- [7] G. Davie, J. Langley, A. Samaranayaka, and M. E. Wetherspoon, "Accuracy of injury coding under ICD-10-AM for New Zealand public hospital discharges," *Injury Prevention*, vol. 14, no. 5, pp. 319–323, oct 2008.
- [8] A. V. Ebbesen, "The creation of the central person registry in denmark," in *IFIP Conference on History of Nordic Computing*. Springer, 2014, pp. 49–57.
- [9] F. Fabris, A. A. Freitas, and J. M. Tullet, "An Extensive Empirical Comparison of Probabilistic Hierarchical Classifiers in Datasets of Ageing-Related Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 6, pp. 1045–1058, jan 2016.
- [10] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, 2016.
- [11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank,

## References

- physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [12] E. R. Hansen, T. Sagi, K. Hose, G. Y. H. Lip, T. B. Larsen, and F. Skjøth, "MIMIC Prescriptions result files," 2020.
- [13] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics*, vol. 216, pp. 574–8, 2015.
- [14] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes," *Computer methods and programs in biomedicine*, vol. 177, pp. 141–153, 2019.
- [15] C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., sep 2017, pp. 3110–3113.
- [16] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [17] F. Li and H. Yu, "Icd coding from clinical text using multi-filter residual convolutional neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8180–8187.
- [18] E. Lynge, J. L. Sandegaard, and M. Rebolj, "The danish national patient register," *Scandinavian journal of public health*, vol. 39, no. 7\_suppl, pp. 30–33, 2011.
- [19] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1614–1623.
- [20] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 1101–1111.
- [21] A. Névél, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of biomedical semantics*, vol. 9, no. 1, pp. 1–13, 2018.
- [22] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: models and evaluation metrics," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 231–237, mar 2014.



## References

- [23] N. Razavian, J. Marcus, and D. A. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, ser. JMLR Workshop and Conference Proceedings, F. Doshi-Velez, J. Fackler, D. C. Kale, B. C. Wallace, and J. Wiens, Eds., vol. 56. JMLR.org, 2016, pp. 73–100.
- [24] A. D. Reys, D. Silva, D. Severo, S. Pedro, M. M. d. S. e Sá, and G. A. Salgado, "Predicting multiple icd-10 codes from brazilian-portuguese clinical notes," in *Brazilian Conference on Intelligent Systems*. Springer, 2020, pp. 566–580.
- [25] M. Ronning, "A historical overview of the atc/DDD methodology," *WHO drug information*, vol. 16, no. 3, p. 233, 2002.
- [26] T. Sagi, E. R. Hansen, K. Hose, G. Y. Lip, T. B. Larsen, and F. Skjøth, "Towards assigning diagnosis codes using medication history," in *International Conference on Artificial Intelligence in Medicine*. Springer, 2020, pp. 203–213.
- [27] M. Schmidt, S. A. J. Schmidt, J. L. Sandegaard, V. Ehrenstein, L. Pedersen, and H. T. Sørensen, "The danish national patient registry: a review of content, data quality, and research potential," *Clinical epidemiology*, vol. 7, pp. 449–490, 2015.
- [28] M. Schmidt, H. T. Sørensen, and L. Pedersen, "Diclofenac use and cardiovascular risks: series of nationwide cohort studies," *bmj*, vol. 362, 2018.
- [29] S. A. Schmidt, M. Vestergaard, L. M. Baggesen, L. Pedersen, H. C. Schønheyder, and H. T. Sørensen, "Prevaccination epidemiology of herpes zoster in denmark: quantification of occurrence and risk factors," *Vaccine*, vol. 35, no. 42, pp. 5589–5596, 2017.
- [30] H. Wallach Kildemoes, H. Toft Sørensen, and J. Hallas, "The danish national prescription registry," *Scandinavian journal of public health*, vol. 39, no. 7\_suppl, pp. 38–41, 2011.
- [31] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, jan 2018.
- [32] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical Multi-Label Classification Networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 5075–5084.
- [33] R. Wockenfuss, T. Frese, K. Herrmann, M. Clausnitzer, and H. Sandholzer, "Three- and four-digit ICD-10 is not a reliable classification system in primary care," *Scandinavian Journal of Primary Health Care*, vol. 27, no. 3, pp. 131–136, jan 2009.
- [34] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on multi-output learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2409–2429, 2019.
- [35] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.



# Paper C

## Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies

Emil Riis Hansen, Tomer Sagi, Katja Hose

The paper has been published in the  
*Proceedings of the 5th International Workshop on Health Data Management in the  
Era of AI (HeDAI@EDBT/ICDT 2023)*, vol 3379, 2023

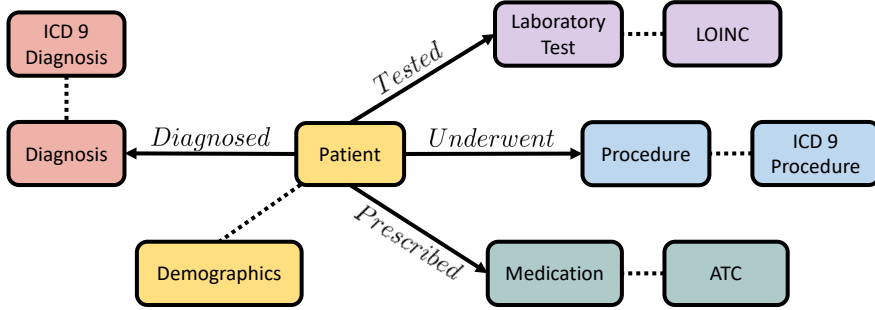
## Abstract

*A variety of hierarchical domain taxonomies exist in the medical domain for describing medical concepts such as laboratory tests, medications, and procedures. The structural information contained within domain taxonomies contains rich semantic information pertaining to the described concepts and their relationships to each other. As AI models are successfully applied in many medical areas, it is only natural to explore integrating AI models with medical domain taxonomies. However, only a few, nascent attempts have been made. In this work, we investigate how the structure of hierarchical medical taxonomies can be used to improve the performance of a diagnosis prediction task. Specifically, we suggest a method titled TreeEmb to pre-initialize the node embeddings of a patient graph derived from electronic health records using information from the taxonomy. We expect this method to improve the performance of graph convolution network models over the enriched patient graph. We evaluate our method over a patient graph created from the MIMIC-IV electronic health record dataset enriched by initializing node embeddings using hierarchical medical taxonomies. We use type-specific domain knowledge from hierarchical medical taxonomies such as the ICD-9 procedures, ATC medication, and LOINC laboratory test taxonomies. Experimental results from a multi-label diagnosis prediction task over this graph demonstrate the efficacy of our approach.*

© The Authors 2023, published under the CC-BY 4.0 License. Reprinted, with permission from Emil Riis Hansen, Tomer Sagi and Katja Hose.

Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies. In: Proceedings of the 5th International Workshop on Health Data Management in the Era of AI (HeDAI@EDBT/ICDT 2023), CEUR Workshop Proceedings, vol 3379, March, 2023. [https://ceur-ws.org/Vol-3379/HeDAI\\_2023\\_paper400.pdf](https://ceur-ws.org/Vol-3379/HeDAI_2023_paper400.pdf)

*The layout has been revised.*



**Fig. C.1:** EHR graph representation relating patients to laboratory tests, procedure codes, and medication intake. Dashed lines represent related information such as patient demographics and hierarchical medical structures such as LOINC, ICD-9 Procedures, and ATC as described in Section 3.2.

## 1 Introduction

The medical domain has accumulated an abundance of domain knowledge structured as hierarchical taxonomies. Integrating semantically rich domain knowledge such as hierarchical taxonomies into Artificial Intelligence (AI) technologies could improve their predictive capabilities in numerous medical applications such as patient diagnosis prediction and protein function prediction using end-to-end supervised learning [25].

Patients’ Electronic Health Records (EHR) can be readily modeled as multi-relational graphs connect patients with their associated medical histories, such as prescriptions, laboratory tests, and procedures, as illustrated in Figure C.1. We, henceforth, name such graphs *EHR graphs*. The AI technology of Graph Convolution Networks (GCNs) has recently become the *de facto* standard for solving many medical problems over EHR graphs due to their seamless ability to learn latent node embeddings for subsequent downstream tasks, such as node classification, link prediction, and whole graph classification in an end-to-end manner [29].

Much work has recently been put into the model-centric development of novel GCN architectures, such as RelationalGCN [24] utilizing the multi-relational nature of graphs and GraphSAGE [9] with a scalable node sampling approach. However, although rich semantic information often exists alongside medical graphs, such as textual descriptions, hierarchical taxonomies, and uncertainty information [14], only a few works investigate in-

corporating such information in a data-centric way for improving classification and regression tasks [4].

As the structure of hierarchical medical domain taxonomies contains human-curated knowledge pertaining to the properties and similarity between taxonomic concepts, we surmise that such structural knowledge can benefit downstream tasks if integrated into AI models. Hence, in this paper, we investigate a method termed *TreeEmb* for encoding the structure of hierarchical medical domain taxonomies to pre-initialize node embeddings in EHR graphs for improved classification performance in a patient diagnosis code prediction task.

This paper is structured as follows; in Section 2, we present related work using domain hierarchies in the initialization of node embeddings and the task of patient diagnosis prediction using graph convolution networks. Section 3 presents the proposed method and theoretical concepts. In Section 4, we present the data used for experimentation, followed by Section 5, where the experimental setup and results are analyzed and explained. Lastly, in Section 6, we conclude and introduce future work.

## 2 Related Work

**Embedding Initialization.** Research into integrating domain information, such as textual descriptions, images, type-hierarchies, and uncertainty information into graph convolution models has lately shown promise [14]. Pre-initializing node embeddings is a central method for integrating auxiliary information with graph convolution networks. Hamilton et al. [9] use text attributes, node profile information, and node degrees to pre-initialize embeddings of three datasets. Zhao et al. [30] use TF/IDF and binary word presence vectors to pre-initialize node embeddings for citation graphs. Other works pre-initialize node embeddings by extracting graphlet features directly from the structure of the input graph [22]. Ali et al. [2] construct manual features such as age and follower count for each social network user. While individual or combinations of manually constructed features have shown promising results for the pre-initialization of node embeddings, none of these works have so far investigated integrating hierarchical domain taxonomies to pre-initialize node embeddings.

**Patient Diagnosis Prediction.** Diagnosis prediction is the vital medical application of finding patient co-morbidities using the patient’s medical history [26]. Hierarchical domain knowledge has recently been introduced into various AI models for diagnosis prediction. In [6], hierarchical medical taxonomies are used to embed medical concepts to leverage the general problem

### 3. Initializing Graph Embeddings

of data insufficiency and model interpretability by learning hierarchical medical concept embeddings, pre-initialized on co-occurrence information by a weighted sum of concept paths. Instead, in this work, we propose using the concept taxonomies for pre-initializing node embeddings of a medical patient graph for subsequent GCN-based diagnosis prediction. The approach by Sun et al. [27] utilizes GCNs on two bipartite graphs, e.g., symptom-relationship and patient-diagnosis, to learn an optimized space wherein patients will have a small distance to assigned diagnosis concepts. However, instead of dividing domain knowledge and patient information into separate bipartite graphs, we investigate the effect of integrating hierarchical auxiliary domain knowledge with a patient graph consisting of multiple patients and their related medical concepts, not limited to symptoms. The work closest to ours is that of [20], in which a knowledge graph is built using auxiliary domain knowledge from the MEDLINE medical corpus for multi-label prediction of patient diseases. Patients are associated with diagnosis codes related to laboratory tests, habits, and profiles in their work. However, different from our work, their method of diagnosis prediction is not related to graph convolutions, and patients are not associated with each other.

## 3 Initializing Graph Embeddings

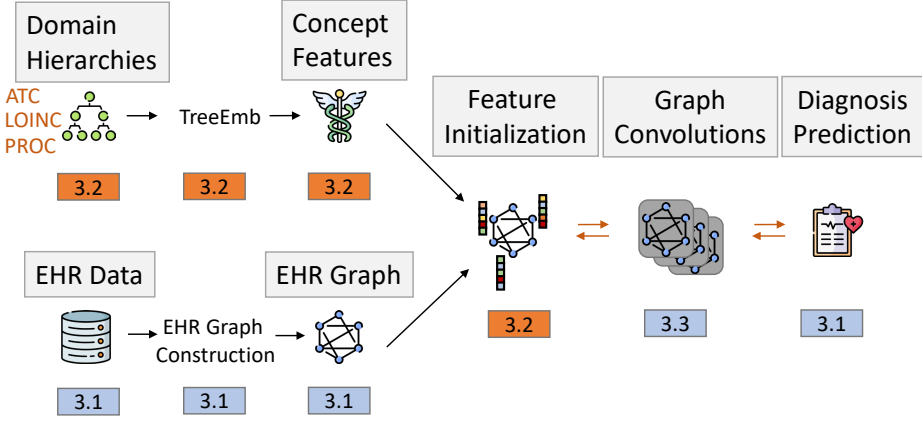
In this section, we formalize our method *TreeEmb* of using hierarchical medical taxonomies to pre-initialize node embeddings for the medical application of multi-label diagnosis prediction. The overall approach is illustrated in Figure C.2, with section references for further details.

An EHR graph is first created from an EHR dataset as detailed in Section 3.1. Concept embeddings are then created from the hierarchical medical taxonomies' structure to derive meaningful latent descriptions of medical concepts and used to pre-initialize node embeddings in the EHR graph as described in Section 3.2. Finally, multiple layers of graph convolutions, as described in Section 3.3, are trained for multi-label patient diagnosis prediction.

### 3.1 Multi-label Diagnosis Prediction over EHR Data

This section introduces how a multi-relational patient-centric graph can be constructed from an EHR dataset and the challenge of multi-label patient diagnosis prediction.

EHR data relate patients to medical concepts such as medications, laboratory tests, and procedures. Given a set of patients  $S$  and a set of medical concepts  $C$ , where  $C^t \subset C$  is the subset of distinct medical concepts types,



**Fig. C.2:** Illustration of the overall approach. Blue boxes reference sections with further details on the specific step. Arrows represent the directional flow of data. Orange boxes represent our primary contribution of pre-initializing graph node embeddings using concept features extracted from hierarchical medical domain taxonomies. Orange arrows describe the parts of the approach that are learned using backpropagation.

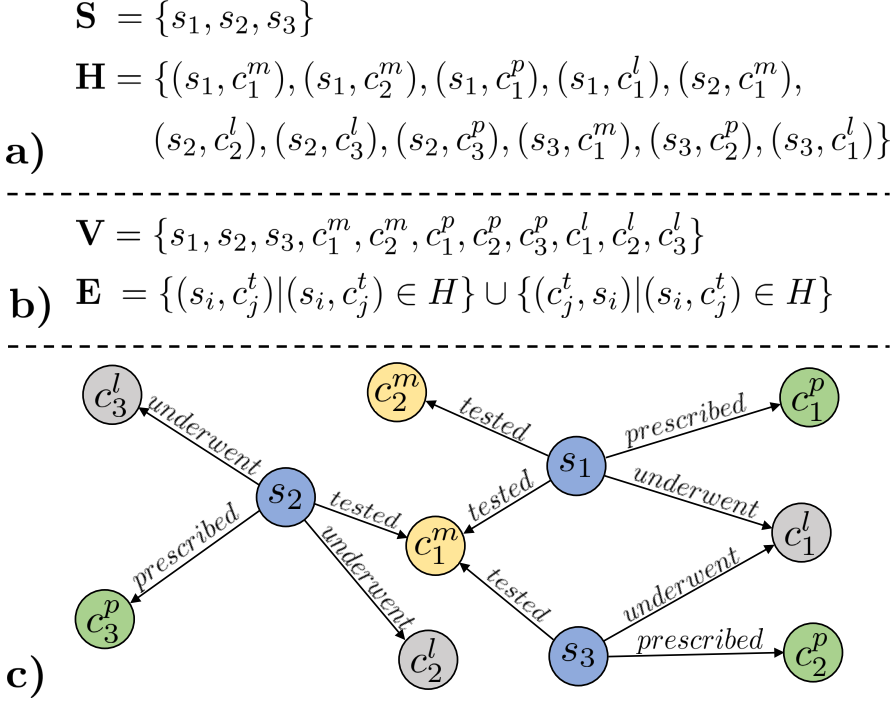
then an EHR dataset can formally be defined as the set  $H$  of tuples  $(s, c)$  relating a patient  $s \in S$  with an associated medical concept  $c \in C$ .

Given the example EHR dataset  $H$  and a set of patients  $S$  as illustrated in Figure C.3a), we create an EHR graph as follows.

The set of graph nodes  $V$  is created as the union between the set of unique patients and the set of unique medical concepts from  $C$  as illustrated in Figure C.3b), and the graph edges are created as the set  $E$  of relations and reversed relations between concepts and patients from  $H$ . Furthermore, every edge in  $E$  is given an edge type as specified by the medical concept type involved in the relation. As an example, the edge  $(s_1, c_1^m)$  could have an edge type of *prescribed* as illustrated in Figure C.3c), as the patient  $s_1$  has been prescribed the medication  $c_1^m$ . The final patient graph created from  $H$  and  $S$  is illustrated in Figure C.3c). For brevity, reverse relations are not depicted in the graph. Over this graph, we define the mapping function  $t_v : V \rightarrow T$  for getting the type of a node, the function  $t_e : E \rightarrow R$  for getting the type  $r$  of an edge, and the function  $f_v : V \rightarrow F$  for getting the embedding  $f_i^t$  of a node  $v_i$  of type  $t = t_v(v_i)$ .

Given an EHR dataset and a set of diagnosis concepts  $D$ , the challenge of patient diagnosis prediction is to find the subset  $D' \subset D$  pertaining to a patient  $s \in S$  s.t.  $D'$  matches the actual set of diagnosis concepts related to the patient. We model this challenge as a multi-label classification problem.

### 3. Initializing Graph Embeddings



**Fig. C.3:** Illustration of the EHR graph creation process. **a)** set of patients  $S$  and an EHR dataset  $H$ . **b)** the graph nodes  $V$  and graph edges  $E$ . **c)** the final graph represented by nodes and typed directed edges. For brevity, reverse relations are not depicted in the graph.

### 3.2 Pre-initialization Using Domain Hierarchies

Node features can be either pre-initialized using entity-specific information or random-initialized and learned as part of the model training process. Pre-initialization of node embeddings can be done by extracting type-specific entity information from the nodes or by extracting features from the graph structure. Examples of the former are pre-trained convolution neural networks for imaging information and natural language processing models for text data. An example of the use of graph structure is by counting substructures such as graphlets [1]. However, an overlooked source of rich semantic information can be found in type-specific domain hierarchies prevalent in many domains. Domain hierarchies are curated hierarchies of related concepts. Inherently, their structure contains knowledge regarding the relationship between concepts, and each hierarchical layer contains information about the properties of its concepts. Hence, we argue that the position of a concept within hierarchies contains rich semantic information.

In the medical domain, structured medical concepts such as medications,

diagnoses, laboratory tests, and procedures are coded in hierarchical taxonomies. Medication can be coded using the world health organization’s anatomical therapeutic classification system (ATC) [21] and classifies medication based on its active ingredients and organ or system. Hence, the location of medications within the hierarchy contains semantic information relevant to the task of diagnosis prediction. As an example, for the medication with code *A10BA02*, e.g., metformin, the first level of the ATC hierarchy specifies that the medication targets the alimentary tract and metabolism system. Level two specifies the therapeutic subgroup, e.g., the drug is used in diabetes. Level three defines the pharmacological subgroup, e.g., the drug lowers blood glucose. The fourth level indicates the chemical subgroup of the drug, in this case, biguanides, and the last level specifies the chemical substance, e.g., metformin. Given that a patient has received metformin, the patient likely suffered from type 2 diabetes. Explicitly integrating such hierarchical information into concept embeddings should enable the AI model to learn from the proximity of similar concepts.

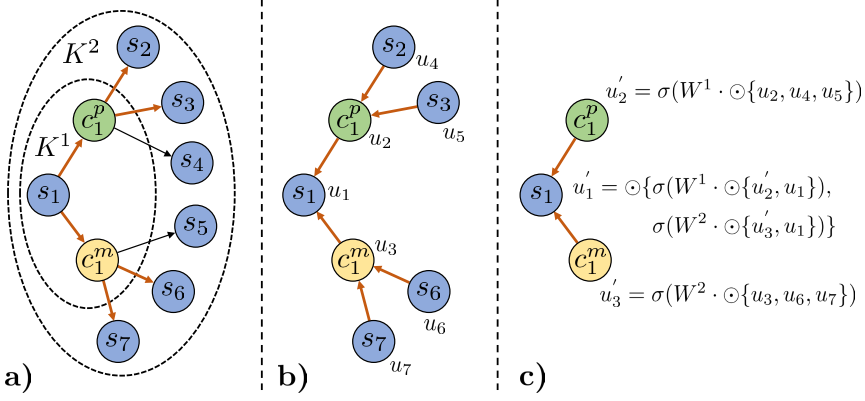
Surgical procedures performed on patients can be coded using the ICD-9 Procedures (PROC) taxonomy [28] grouping related procedures based on their site of operation. Given that a patient has received the surgical procedure with code 07.2, e.g., partial adrenalectomy, the patient likely suffered from a disease related to the endocrine glands.

Laboratory tests can be coded using the LOINC concept codes [18] over which a hierarchical taxonomy exists, grouping related laboratory tests by their class, component, and system, providing valuable information on the purpose of laboratory tests.

Using the aforementioned hierarchical medical taxonomies, and the example of the medical concept with code *B02AA02*, e.g., Tranexamic acid from the ATC hierarchy, we propose the *TreeEmb* method for pre-initializing node embeddings using type-specific hierarchical domain knowledge. Starting from 0, a unique index is assigned to each node in the tree as illustrated in Figure C.5b). Subsequently, a depth-first search is performed from the root of the hierarchical domain taxonomy to each leaf node for collecting the indexes along the shortest path to each leaf. Suppose the concept *B02AA02* is given the initial index 3, then the indexes between *B02AA02* and the root node is  $[0, 1, 3]$  as illustrated in Figure C.5c). Eventually, leaf nodes are assigned an embedding as the one-hot encoded version of their shortest path indexes. As the concept *B02AA02* has accumulated indexes  $[0, 1, 3]$  and as the example tree has 11 nodes, *B02AA02* is assigned an embedding vector of dimensionality 11 with 1 in the positions 0, 1, and 3 and 0 in every other position as illustrated in Figure C.5d). The computed features of tree leaf concepts can then be used to pre-initialize node embeddings. Furthermore, using this embedding technique ensures that concepts closely related in the tree will have similar embeddings compared to concepts far away. Hence, we



### 3. Initializing Graph Embeddings

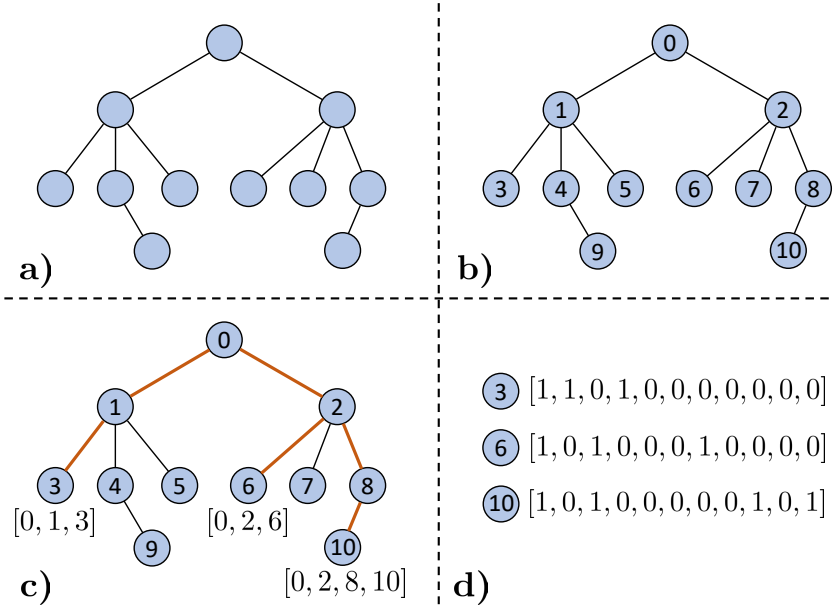


**Fig. C.4:** Example steps of our graph convolution. **a)** A 2-layer 2-node fanout sampling strategy finds the neighborhood  $K^1$  of patient  $s_1$ . Each of the sampled nodes  $\{c_1^p, c_1^m\}$  uses the same sampling strategy on their immediate neighborhood  $K^2$  to further sample nodes  $\{s_2, s_3, s_6, s_7\}$ . **b)** The sampled subgraph with node features  $u_1$  to  $u_7$  as extracted through the node feature mapping  $f_v$ . **c)** Our combined graph convolution aggregate and update step. A relation-specific transformation matrix  $W^i$  is applied to the element-wise mean  $\odot$  of similar typed entities as done in [9]. Finally, a non-linear activation function  $\sigma$  is applied to individual convolutions. If different typed features are to be combined as in the combination of  $\{u_1, u_2, u_3\}$  the element-wise mean combines individual transformations.

conjecture that GCNs will be more easily able to learn that groups of closely related concepts are used in treating the same disease, thus decreasing the epistemic uncertainty by adding domain knowledge.

### 3.3 Graph Convolution Networks

Graph convolutions can learn from the structure of graphs by propagating node features between neighboring nodes using learnable *aggregation* and *update* functions as illustrated in Figure C.4. Aggregation functions combine neighborhood information by imposing transformation matrices on the output of the neighborhood aggregation. Update functions, then learn how to integrate information from the current node embedding and the features of the neighborhood aggregation function. We employ a multi-relational variant of the GraphSAGE [9] algorithm for learning latent node embeddings for graphs with multiple relation types between concept nodes by exploiting not only the structure but also the multi-relational nature of EHR graphs.



**Fig. C.5:** *TreeEmb* method for constructing concept embeddings from hierarchical taxonomies. **a** A tree-structured hierarchical taxonomy. **b** breadth first search indexes every node from 0. **c** Depth-first search from the root to each leaf node collects shortest path indexes. **d** One-hot encoding generates embeddings for leaf node concepts.

## 4 Data

We perform experiments on the MIMIC-IV [15] EHR dataset from PhysioNet [8] consisting of 382,278 intensive care unit patients from the Beth Israel Deaconess Medical Center from the period 2008 to 2019. MIMIC-IV encompasses laboratory results, vital signs, diagnoses ascertained, administered medications, and demographics. The data is structured as a relational database.

To disambiguate medical concepts, we transform the dataset into the observational medical outcomes partnership (OMOP) common data model (CDM) [13] using an extract-transform-load (ETL) conversion flow.<sup>1</sup> The CDM format disambiguates and standardizes medical concepts and thus provides a means of interoperability for subsequent AI models to operate on disparate medical datasets converted into the CDM. In the CDM format laboratory tests are coded using the LOINC taxonomy, procedures are coded using the ICD-9 procedures taxonomy, and laboratory tests are coded using

<sup>1</sup><https://github.com/OHDSI/MIMIC>

#### 4. Data

the RxNorm taxonomy [19]. Since RxNorm is a flat taxonomy, we map each medication concept through its active ingredients to the hierarchical ATC medication taxonomy.

**Table C.1:** Number of distinct concepts for EHR data types.

Data Type	Distinct Concepts
Medication	1,749
Diagnosis	537
Laboratory Test	1,328
Procedure	1,228

For patient multi-label diagnosis prediction, we build the EHR graph based on patient diagnostic EHR concept types used in related work in EHR-based diagnosis prediction [12, 17, 26] and end up with demographic information, prescriptions, procedures, laboratory tests, and the task labels as patient diagnosis codes.

Patient diagnosis codes are coded using the 9th version of the International Classification of Diseases (ICD-9) and consist of approximately 13,000 diagnosis codes [5]. We omit codes related to the ICD-9 E and V hierarchies as these are related to external causes of injury and are generally not discernible by EHR data. We further omit hierarchies of codes as summarized in Table C.2. Omitting these hierarchies, we are left with 8,681 disease codes. Since it is usually not possible to generalize from a low number of cases, we omit codes for which less than 500 patient cases exist. We are ultimately left with 128,605 patients diagnosed with a total of 1,054,670 diagnoses from 537 distinct diagnosis codes. The full list of 537 diagnosis codes are available online<sup>2</sup>. Table C.1 summarizes the number of distinct concepts for each medical EHR concept type.

**Table C.2:** Summarizing disease codes omitted from further analysis.

Codes	Count	Description
290 – 319	375	Mental Disorders
630 – 679	530	Comp. of Pregnancy
780 – 799	330	Injuries and Poison
800 – 999	1,617	Ill-Defined Conditions
E and V	1,467	Ext. Causes of Injury

---

<sup>2</sup>[https://github.com/dkw-aau/graph\\_embedding\\_initialization](https://github.com/dkw-aau/graph_embedding_initialization)

## 5 Experiments and Results

To investigate the effect of pre-initializing node embeddings using domain hierarchies, we conduct several empirical experiments as summarized in Table C.3 using the model pipeline as illustrated in Figure C.2. Each experiment is trained on the problem of multi-label patient diagnosis prediction using a multi-relational version of the GraphSAGE algorithm as described in Section 3 with the input EHR dataset described in Section 4. In the **Rand** experimental setting, initial graph node embeddings are random-initialized using Xavier initialization [7] and made trainable as part of the supervised model training phase [10]. Hence, **Rand** serves as a transductive baseline experiment. Transductive methods generally perform better on subsequent downstream prediction tasks, however, with the cost of not being able to extrapolate to unseen examples [9].

Table C.3: Overview of experimental settings.

Experiment	Learning	Embedding Data
<b>FeatInit</b>	Inductive	Hierarchical Taxonomies
<b>Rand</b>	Transductive	Xavier Initialization
<b>Graphlet</b>	Inductive	Graph Structure

In the **Graphlet** experimental setting, features are pre-initialized using state-of-the-art graphlet and edge count features [1] as in [22]. **Graphlet** serves as an inductive baseline experiment, as trained models can extrapolate to unseen examples.

The **FeatInit** experimental setting investigates the effect of pre-initializing node concept embeddings using the latent information contained within hierarchical medical taxonomies using the *TreeEmb* method as described in Section 3. In **FeatInit**, node embeddings should already contain domain information relevant to the task of diagnosis prediction; hence embeddings are kept constant during training. Furthermore, in the **FeatInit** experimental setting, patient features are pre-initialized using categorical values for sex, race, and ethnicity and a continuous variable for the patient’s age. Moreover, as **FeatInit** does not train node embeddings, trained models can extrapolate to unseen examples.

### 5.1 Experimental Details

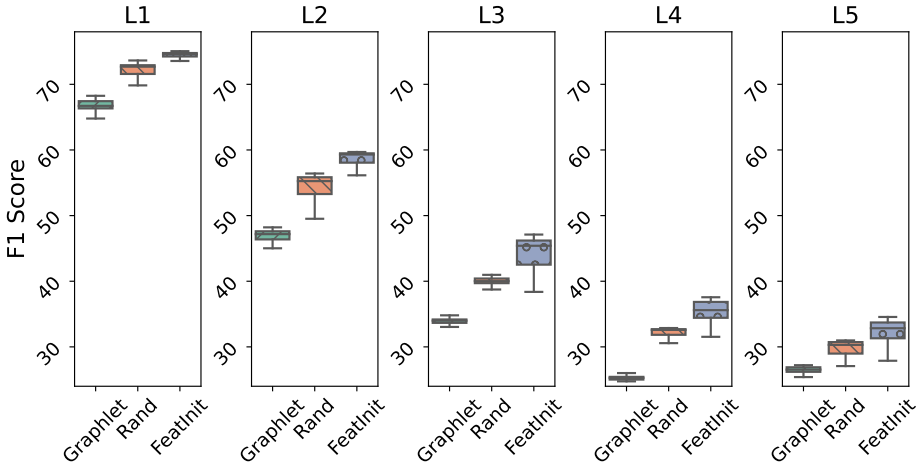
For each experiment, we perform 100 iterations of tree-based Parzen estimation (TPE) [3] for hyperparameter optimization over the set of parameters as summarized in Table C.4. Each iteration is trained using the Adam [16] variation of stochastic gradient descent with binary cross-entropy as the loss func-

## 5. Experiments and Results

**Table C.4:** Parameter settings for hyperparameter optimization using tree-based Parzen estimation. *U* means uniform distribution.

Parameter	Values
Model Depth	{2, 3}
Learning Rate	{1e-3, 5e-3, 1e-2}
Dropout	$U(0.0..0.5)$
Hidden Dim	{32, 64, 128, 256}

tion. Each experimental setting is investigated on the prediction of five sets of diagnosis codes as in [11, 23], with each set relating to a level of aggregation on the hierarchical ICD-9 diagnosis taxonomy. In the first setting, named *L5*, the task is to predict the raw comorbidities of patients from the entirety of the 537 diagnosis codes as described in Section 4. The remaining settings investigate diagnosis code prediction on aggregated levels of the ICD-9 diagnosis taxonomy named *L4* through *L1* with 427 disparate diagnosis codes for *L4* to 13 disparate diagnosis codes for *L1*. Aggregating diagnosis codes enables us to investigate the effect of pre-initializing graph concept embeddings from hierarchical medical taxonomies extracted through *TreeEmb* on classification problems of varying complexities.



**Fig. C.6:** Experimental results of diagnosis code prediction on five sets of diagnosis codes for the experimental settings **Rand**, **Graphlet**, and **FeatInit**.

As graph convolutions require the same dimensionality for each node type, we do an initial transformation on node input features using type-specific non-linear transformations into the feature dimensionality required by the graph convolution layers. Thus, the transformation is learned end-

**Table C.5:** Experimental results in terms of harmonic mean F1 scores for the experimental settings **Graphlet**, **Rand**, and **FeatInit** on five diagnosis code prediction problems with varying number of classes. Imp. presents the relative improvement in terms of F1 value for initializing concept embeddings using the *TreeEmb* embeddings.

Setting	Median			Imp.
	Graphlet	Rand	FeatInit	
L5 - 537 codes	26.54	30.30	32.84	<b>2.54</b>
L4 - 427 codes	25.27	32.58	35.60	<b>3.02</b>
L3 - 229 codes	34.00	40.01	45.40	<b>5.39</b>
L2 - 61 codes	47.18	55.25	59.30	<b>4.05</b>
L1 - 13 codes	66.72	72.69	74.58	<b>1.89</b>
Setting	Best			Imp.
	Graphlet	Rand	FeatInit	
L5 - 537 codes	27.21	30.98	34.56	<b>3.58</b>
L4 - 427 codes	26.01	32.87	37.56	<b>4.69</b>
L3 - 229 codes	34.81	40.97	47.11	<b>6.14</b>
L2 - 61 codes	48.21	56.41	59.69	<b>3.28</b>
L1 - 13 codes	68.25	73.63	75.05	<b>1.42</b>

to-end with the task of diagnosis prediction. Additionally, we transform the output node embeddings as computed by the final convolution layer using a non-linear transformation into the dimensionality of the number of diagnosis codes in a specific level of ICD-9 aggregation, such that we end up with one output node for each predictable diagnosis code. We split patients into training validation and test sets with sizes 80/10/10 and used early stopping based on validation loss.

To evaluate and compare across experimental settings, we use the standard harmonic mean F1 value between the micro-averaged precision and recall as it is commonly used in the evaluation of multi-label classification tasks [20]. Furthermore, to investigate the robustness of pre-initializing features using *TreeEmb* embeddings, we evaluate the median over all 100 model iterations for each experiment. All experimental code and data are available online<sup>3</sup>.

## 5.2 Results and Analysis

Figure C.6 presents the results for each experimental setting over all iterations of the TPE. Experimental results in terms of the F1 value for the median and best-performing models are summarized in Table C.5.

As illustrated in Figure C.6, using *TreeEmb* embeddings for pre-initializing node features resulted in improved F1 scores compared to learning node em-

<sup>3</sup>[https://github.com/dkw-aau/graph\\_embedding\\_initialization](https://github.com/dkw-aau/graph_embedding_initialization)

## 5. Experiments and Results

beddings as part of the training and pre-initialization using graphlet features. Furthermore, using unpaired t-test between **Rand** and **FeatInit** and between **Graphlet** and **FeatInit** results for any level of diagnosis code aggregation results in the two-tailed P value  $p < .001$ , which by conventional criteria indicates a statistically significant difference between the two groups.

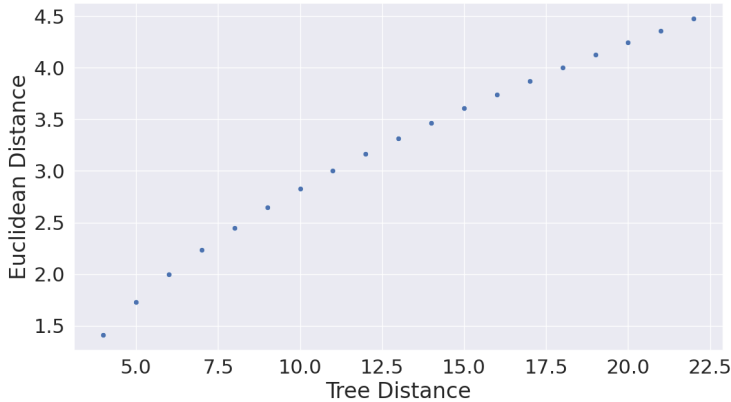


Fig. C.7: Monotonicity of LOINC concept embedding space.

As summarized in Table C.5, for each setting, the best performing **FeatInit** model outperforms the best performing **Rand** and **Graphlet** models by 1.42 – 6.14 and 6.80 – 12.30 percentage points in terms of F1 score respectively. These results indicate that the initialization of node features using the hierarchical knowledge contained within domain taxonomies could provide valuable knowledge for solving domain-specific problems such as the medical problem of patient diagnosis prediction.

The embeddings produced by *TreeEmb* should reflect the structure of the hierarchical taxonomy. Assuming that semantically similar concepts are close in the tree and disparate concepts far from each other, the distance between constructed embeddings should increase as the path length between nodes in the tree increases. To investigate this aspect of the *TreeEmb* embeddings, we compared the Euclidean distance between pairs of concept embeddings with the length of the shortest path on the tree between the pairs. As illustrated in Figure C.7, the Euclidean distance between node embeddings is a monotonic increasing function given the length of the shortest path between nodes. This means that similar concepts will have similar embeddings while dissimilar concepts will have disparate embeddings.

## 6 Conclusion

In this work, we proposed that hierarchical medical taxonomies contain valuable knowledge that can be utilized by the pre-initialization of graph node embeddings. We then presented a method termed *TreeEmb* to do so. We evaluated the proposed method on the medical problem of multi-label diagnosis prediction by constructing *TreeEmb* embeddings for the pre-initialization of concept nodes in an EHR graph for the three medical hierarchical taxonomies ATC, LOINC, and ICD-9 Procedures. Experimental results from the prediction task on five different sets of diagnosis codes of varying difficulty demonstrate the superiority of *TreeEmb* embeddings over a transductive baseline of learned concept embeddings and an inductive baseline of pre-computed graphlet features. All experimental code and data are available online<sup>4</sup>.

For future work, we aim to investigate the proposed method in domains beyond the medical. Furthermore, since not all levels of hierarchical domain taxonomies may be equally important for the given prediction task, we aim to investigate trainable attention mechanisms for constructing concept embeddings from only the most relevant hierarchical knowledge. We also aim to explore other graph convolution models, including attention techniques.

## References

- [1] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield, “Efficient graphlet counting for large networks,” in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 1–10.
- [2] S. A. Alhosseini, R. B. Tareaf, P. Najafi, and C. Meinel, “Detect me if you can: Spam bot detection using inductive representation learning,” in *WWW (Companion Volume)*. ACM, 2019, pp. 148–153.
- [3] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” in *NIPS*, 2011, pp. 2546–2554.
- [4] J. D. Bossér, E. Sörstadius, and M. H. Chehreghani, “Model-centric and data-centric aspects of active learning for deep neural networks,” in *IEEE BigData*. IEEE, 2021, pp. 5053–5062.
- [5] D. J. Cartwright, “Icd-9-cm to icd-10-cm codes: what? why? how?” in *Advances in wound care*. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, 2013, pp. 588–592.
- [6] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, “Gram: graph-based attention model for healthcare representation learning,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 787–795.

---

<sup>4</sup>[https://github.com/dkw-aau/graph\\_embedding\\_initialization](https://github.com/dkw-aau/graph_embedding_initialization)



## References

- [7] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
- [8] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, and et al., "Physiobank, physiotookit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [9] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.
- [10] —, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.
- [11] E. R. Hansen, T. Sagi, K. Hose, G. Y. Lip, T. B. Larsen, and F. Skjøth, "Assigning diagnosis codes using medication history," *Artificial Intelligence in Medicine*, p. 102307, 2022.
- [12] A. Hosseini, T. Chen, W. Wu, Y. Sun, and M. Sarrafzadeh, "Heteromed: Heterogeneous information network for medical diagnosis," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 763–772.
- [13] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, and et al., "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics*, vol. 216, pp. 574–8, 2015.
- [14] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.
- [15] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. Celi, and R. Mark, "Mimic-iv (version 0.4). physionet," 2020.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.
- [17] Z. Liu, X. Li, H. Peng, L. He, and S. Y. Philip, "Heterogeneous similarity graph neural network on electronic health records," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1196–1205.
- [18] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, and et al., "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.
- [19] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore, "Normalized names for clinical drugs: Rxnorm at 6 years," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 441–448, 2011.
- [20] T. Pham, X. Tao, J. Zhang, J. Yong, Y. Li, and H. Xie, "Graph-based multi-label disease prediction model learning from medical data and domain knowledge," *Knowledge-Based Systems*, p. 107662, 2021.

## References

- [21] M. Ronning, "A historical overview of the atc/DDD methodology," *WHO drug information*, vol. 16, no. 3, p. 233, 2002.
- [22] R. A. Rossi, R. Zhou, and N. K. Ahmed, "Deep inductive network representation learning," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 953–960.
- [23] T. Sagi, E. R. Hansen, K. Hose, G. Y. Lip, T. B. Larsen, and F. Skjøth, "Towards assigning diagnosis codes using medication history," in *International Conference on Artificial Intelligence in Medicine*. Springer, 2020, pp. 203–213.
- [24] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.
- [25] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging technology in modelling and graphics*. Springer, 2020, pp. 99–111.
- [26] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts, "Deep representation learning of patient data from electronic health records (ehr): A systematic review," *Journal of Biomedical Informatics*, vol. 115, p. 103671, 2021.
- [27] Z. Sun, H. Yin, H. Chen, T. Chen, L. Cui, and F. Yang, "Disease prediction via graph neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 818–826, 2020.
- [28] WHO et al., *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization, 1978.
- [29] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.
- [30] Z. Zhao, H. Zhou, L. Qi, L. Chang, and M. Zhou, "Inductive representation learning via cnn for partially-unseen attributed networks," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 695–706, 2021.

# Paper D

## Patient Event Sequences for Predicting Hospitalization Length of Stay

Emil Riis Hansen, Thomas Dyhre Nielsen, Thomas Mulvad,  
Mads Nibe Strausholm, Tomer Sagi, Katja Hose

The paper has been published in the  
*Proceedings of the 21st International Conference of Artificial Intelligence in  
Medicine (AIME 2023)*, pp. 51-56, 2023  
DOI: [10.1007/978-3-031-34344-5\\_7](https://doi.org/10.1007/978-3-031-34344-5_7)

## Abstract

*Predicting patients' hospital length of stay (LOS) is essential for improving resource allocation and supporting decision-making in healthcare organizations. This paper proposes a novel transformer-based model, termed Medic-BERT (M-BERT), for predicting LOS by modeling patient information as sequences of events. We performed empirical experiments on a cohort of 48k emergency care patients from a large Danish hospital. Experimental results show that M-BERT can achieve high accuracy on a variety of LOS problems and outperforms traditional non-sequence-based machine learning approaches.*

© The Authors 2023. Reprinted, with permission from Emil Riis Hansen, Thomas Dyhre Nielsen, Thomas Mulvad, Mads Nibe Strausholm, Tomer Sagi, and Katja Hose.

Patient Event Sequences for Predicting Hospitalization Length of Stay. In: Proceedings of the 21st International Conference of Artificial Intelligence in Medicine (AIME 2023), Lecture Notes in Computer Science, volume 13897, pages 51-56, June, 2023. [https://doi.org/10.1007/978-3-031-34344-5\\_7](https://doi.org/10.1007/978-3-031-34344-5_7)

*The layout has been revised.*

# 1 Introduction

Increasingly scarce hospital resources challenge (often oversaturated) hospital wards, negatively impacting the quality of health care [1]. Models for predicting the remaining patient hospitalization time, i.e., patient length of stay (LOS), is a valuable tool for healthcare facilities in resource availability planning of, e.g., beds and staff. For instance, prediction of discharge time can be used to preemptively free in-hospital resources to alleviate hospital ward oversaturation [11]. However, LOS prediction is a challenging problem, requiring methods for handling missing data [6] and temporal event dependencies integration.

Previous work on LOS prediction often models patient hospitalizations using standard ML models, such as RFs, GBs, and ANNs, relying on imputation techniques for replacing missing values [2]. However, missing observations in healthcare data are often not missing at random (NMAR), and the mere fact that observations are missing is essential information [6].

To alleviate this and other shortcomings, attention-based models have recently been investigated for sequence structured Electronic Health Record (EHR) data [8, 10]. Attention models address the inefficiency of recurrent networks for long sequences [10] while still capturing significant sequential information by learning from the order of tokens in a sequence. However, in medical data, observations are often grouped with the same timestamp. For example, a blood panel drawn from a patient contains several measurements whose order is undefined. Based on layers of transformer encoders, we propose the Medic-BERT (M-BERT) model inspired by the original BERT model [3]. Using sequences of in-hospital medical events exhibiting event concurrences common in EHR data, we employ M-BERT for LOS prediction. We evaluate M-BERT on a cohort of 48k patient admissions from a large Danish hospital with information on diverse medical events, such as measurements of vital parameters, medication administration, laboratory tests, and conducted procedures.

## 2 Transformer Models for EHR

Patient hospitalizations can naturally be modeled as sequences of medical codes for determining, measuring, or diagnosing the patients' conditions. To standardize how procedures are described, medical facilities code concepts using accepted taxonomies. Hence, hospitalization can be described as a sequence of concept tokens detailing the medical procedures pertaining to a patient and coded using taxonomical concepts. For some procedures, such as vitals and lab tests, a numerical measurement value accompanies the procedure. These measurement values are mapped into normal, abnormal-low,

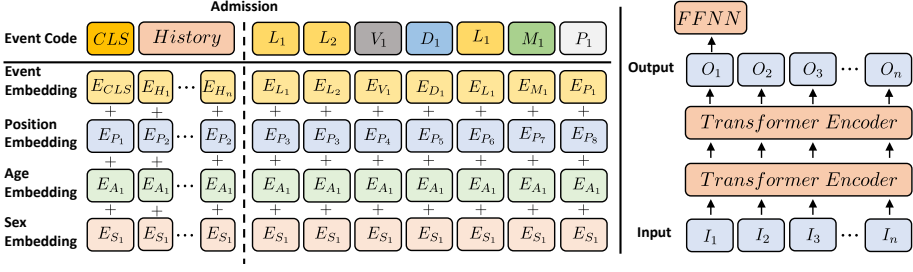


Fig. D.1: Medical event sequence pre-pended with the patient’s medical history.

or abnormal-high tokens.

Furthermore, as patient history is essential for correct treatment, we pre-pend the patient’s medical history to the hospitalization sequence as a tokenized vector consisting of 38 tokens. These include comorbidities (Charlson Index [12]), five years of prescription history grouped by the first level of the ATC hierarchy [9], and the mode, time, and initial hospitalization triage category [13].

In this paper, we propose Medic-BERT (M-BERT), see Fig. D.1, an EHR-data modification of the Bidirectional Encoder Representations from Transformers (BERT) model [3]. BERT is an NLP model based on a stack of encoder layers. The transformer encoder naturally handles complex long-term dependencies that occur between medical concepts through its utilization of multi-head self-attention. Furthermore, BERT naturally operates in domains with irregular intervals between events, as is the case with EHR data. BERT can also naturally integrate disparate input, such as diagnostic and therapeutic events, encoding each event as an n-dimensional vector.

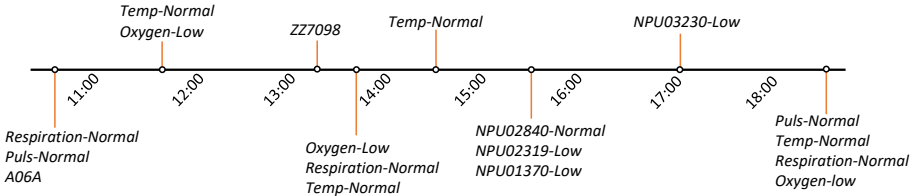


Fig. D.2: Patient event sequence illustrating events grouping together.

M-BERT learns embeddings for the demographic features and the medical event tokens. Together with positional embeddings, the model can learn temporal dependencies within a sequence. We use a static positional embedding modified for usage on medical event sequences: medical events with no natural chronological ordering (e.g., multiple lab tests done on the same sample as illustrated in Fig. D.2) are assigned the same positional embedding.

### 3. Empirical Evaluation and Results

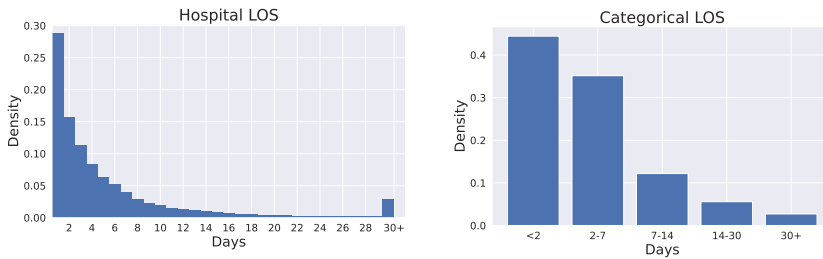


Fig. D.3: Length of stay distribution over patient hospitalizations.

Finally, we use a classification (CLS) token as a final aggregate representation fed to a linear output layer for LOS classification and regression tasks.

## 3 Empirical Evaluation and Results

### 3.1 Data and Experimental Setting

Table D.1: Concept types with occurrences.

Event Type	#Tokens	#Events
Lab Tests	748	2,774,790
Vitals	22	837,931
Medication	1,441	376,591
Procedures	2049	247,924
History	81	1,880,580

We compiled a dataset of hospital emergency care admissions in northern Jutland (Denmark) between 2018-2021. The dataset comprises 48,177 admissions (>one day). Fig. D.3 presents the remaining length of stay distribution. Tab. D.1 summarizes the event types present in the data.

Due to our focus on a single prediction task, we directly train model parameters and token embeddings toward the downstream task of LOS prediction without any unsupervised pre-training.<sup>1</sup> Using the data gathered within 24 hours of admission, we evaluate M-BERT on three LOS prediction tasks: **Binary** classification of  $\text{LOS} > 2$  days, a three-class **Category** task of  $\text{LOS} > 2$ ,  $2 \leq \text{LOS} \leq 7$ , and  $\text{LOS} > 7$  days with class balances as shown in Fig. D.3, and the **Real** regression task of predicting LOS in days (decimal).

We compare our approach with an RF, ANN, and SVM model as implemented in the Sklearn library [7] using default hyper-parameters. The last

<sup>1</sup>[https://github.com/dkw-aau/medic\\_transformer](https://github.com/dkw-aau/medic_transformer)

measured value for each event type (within 24 hours) defined the input features [5]. For features with missing values, we relied on imputation using the mean of the features, and subsequently scaled the values to be between 0 and 1. A  $\chi^2$  test was finally used for extracting the 50 most relevant features.

M-BERT was trained with a  $1e-5$  learning rate on a 80/10/10 random split of patients. We use the evaluation loss for early stopping within ten epochs. The model architecture has six hidden layers with an intermediate layer size of 288, eight attention heads, and an input token embedding size of 288. We truncate sequences to 256 tokens, as most sequences adhere to this limit. To counter overfitting, we add a 10% dropout layer after the final encoder layer, a 10% attention dropout, and a weight decay of 0.003. Further details are available in the extended version of this paper [4].

### 3.2 Results

Table D.2 presents AUROC and F1 scores for the **Binary** and **Category** experiments and Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the regression task. Model performance is stable for different age groups. The results indicate that M-BERT can leverage temporal dependencies inherent to EHR data for increased predictive accuracy. Being a transformer-based model, M-BERT overcomes the challenge of missing data as patient sequences are not required to contain the same events or be of the same length.

Table D.2: Experimental results.

	Binary		Category		Real	
	AUROC	F1	AUROC	F1	MAE	MSE
RF	0.72	0.70	0.66	0.45	4.18	39.08
ANN	0.67	0.68	0.63	0.43	4.09	38.10
SVM	0.70	0.70	0.65	0.38	3.56	43.36
M-BERT	<b>0.78</b>	<b>0.77</b>	<b>0.74</b>	<b>0.54</b>	<b>3.42</b>	<b>37.48</b>

## 4 Conclusion

We have proposed a novel approach for predicting LOS by modeling patient information as event sequences. We adapt the transformer-based architecture to sequence prediction over grouped events of varying data types as typically found in medical event sequences. Our empirical evaluation on a large cohort of emergency care patients from a Danish hospital demonstrates high accuracy on various LOS problems, while also outperforming traditional non-sequence-based approaches. Future work includes model pre-training as well as evaluation of the predictive uncertainty offer by the model. Overall,



the proposed approach has the potential to improve resource allocation in healthcare organizations by providing accurate and reliable predictions of LOS.

## References

- [1] B. af Ugglas, T. Djärv, P. L. Ljungman, and M. J. Holzmann, "Association between hospital bed occupancy and outcomes in emergency care: a cohort study in stockholm region, sweden, 2012 to 2016," *Ann. Emerg. Med.*, vol. 76, no. 2, pp. 179–190, 2020.
- [2] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar, "Machine learning in the prediction of medical inpatient length of stay," *Intern. Med. J.*, vol. 52, no. 2, pp. 176–185, 2022.
- [3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT '19*, 2019, pp. 4171–4186.
- [4] E. R. Hansen, T. D. Nielsen, T. Mulvad, M. N. Strausholm, T. Sagi, and K. Hose, "Hospitalization length of stay prediction using patient event sequences," 2023.
- [5] S. Iwase, T.-a. Nakada, T. Shimada, T. Oami, T. Shimazui, N. Takahashi, and et al., "Prediction algorithm for icu mortality and length of stay using machine learning," *Scientific reports*, vol. 12, no. 1, pp. 1–9, 2022.
- [6] J. Li, X. S. Yan, D. Chaudhary, V. Avula, S. Mudiganti, H. Husby, and et al., "Imputation of missing values for electronic health record laboratory data," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–14, 2021.
- [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. e. a. Grisel, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.
- [8] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.
- [9] M. Ronning, "A historical overview of the atc/DDD methodology," *WHO drug information*, vol. 16, no. 3, p. 233, 2002.
- [10] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *AAAI '18*, vol. 32, 2018.
- [11] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework," *PLOS Digital Health*, vol. 1, no. 4, p. e0000017, 2022.
- [12] V. Sundararajan, T. Henderson, C. Perry, A. Muggivan, H. Quan, and W. A. Ghali, "New icd-10 version of the charlson comorbidity index predicted in-hospital mortality," *Journal of clinical epidemiology*, vol. 57, no. 12, pp. 1288–1294, 2004.
- [13] S. C. Wireklint, C. Elmqvist, and K. E. Göransson, "An updated national survey of triage and triage related work in sweden: a cross-sectional descriptive and comparative study," *SJTREM '21*, vol. 29, no. 1, pp. 1–8, 2021.

## References

# Paper E

## Multi-modal Representation Learning for Medical Analytics

Emil Riis Hansen, Tomer Sagi, Katja Hose

*Under Review*

## Abstract

*Machine learning-based analytics over uni-modal medical data has shown considerable promise and is rapidly being deployed in routine diagnostic procedures. Patient data is diverse and comprised of multiple data types. Multi-modal approaches promise to revolutionize our ability to provide personalized care. Nascent attempts to combine two modalities in a single diagnostic task have utilized the evolving field of multi-modal representation learning (MRL), which learns a shared latent space between related modality samples. This new space can be used to improve the performance of machine-learning-based analytics. However, our understanding of how modalities have been applied in MRL-based medical applications and what modalities are best suited for which medical tasks is still unclear. In this work, we explore the landscape of MRL for medical tasks by presenting a framework for positioning MRL techniques and medical modalities to highlight opportunities for advancing medical applications. We demonstrate our approach by reviewing and classifying more than 1000 papers related to medical analytics using our proposed framework in the most extensive review of its kind to date. We further provide an online tool for researchers and developers of medical analytics to dive into the rapidly changing landscape of MRL for medical applications.*

*The layout has been revised.*

# 1 Introduction

The world is inherently multi-modal. Entities, from patients to proteins, can be described in various ways called modalities. The onset of various diseases and conditions can be measured in a medical setting through a measurable change in biomarker modalities, such as blood pressure, heart rate, and x-ray findings. As an example, the progression of Alzheimer’s Disease (AD) has shown correlation with modalities such as Magnetic Resonance Imaging (MRI) [20], Positron Emission Tomography (PET) [8], and protein measures of Cerebrospinal Fluid (CSF) [4]. MRI provides a means of detecting atrophied brain regions, PET can reveal hypometabolism [27], and protein measures of CSF can detect the presence of beta-amyloid ( $A\beta_{42}$ ), and tau ( $\tau$ ) proteins characteristic of AD [37]. Each modality provides unique information, which combined, could be used to perform medical analytic tasks such as AD progression classification.

Medical machine learning (ML)-based analytics attempt to improve the quality and speed of previously manual tasks and have featured predominantly uni-modal approaches. Combining multiple information modalities in the same manner as a physician considers multiple sources of information can enhance the performance of complex predictive ML-based analytics.

Multi-modal representation learning (MRL) [3]) is a theoretical and practical framework for combining multiple information modalities to improve the effectiveness of ML-based tasks ranging from video classification to emotion recognition. MRL has recently expanded into the medical analytics domain, where it has been used to combine multiple medical modalities for diagnosis and prognosis tasks [6]. However, no comprehensive survey of MRL in the medical domain has been performed, leaving researchers to piece together which combinations of medical information modalities have been attempted for various medical analytics using disparate MRL techniques. Furthermore, various medical information modalities such as omics data, medical images, textual medical records, electronic health records (EHR), computerized clinical practice guidelines, and biomedical knowledge graphs exist in the medical space. It has become a daunting task to sift through these options with medical analytics in mind and identify which are relevant, has been used previously, and in what combination of modalities.

In this work, we investigate MRL as a technique for utilizing multiple sources of information to improve the performance of ML-based medical analytics. We describe and classify the different MRL techniques. We provide a hierarchy of medical information modalities over which one could attempt MRL. We do a comprehensive, structured survey of publications utilizing MRL for ML-based medical analytics, positioning them in the MRL classification space, the medical information modality classification space, a medical

application classification space, and a utility classification space. The MRL classification space describes the specific MRL technique used. The medical information modality classification space describes the modalities being integrated. The medical application classification space describes the clinical motivation, and the utility classification space describes the intended use of the medical analytic. Finally, we create an online electronic business intelligence (BI) tool to demonstrate how these spaces can be used to understand which medical information modalities have been used together and what MRL techniques are appropriate for specific medical analytics.

Our contribution can be summarized as follows. We provide a comprehensive review of MRL technologies that extends upon previous surveys with a novel and updated MRL classification space, including neural network technologies as a top-level category. We further expand this category with recent technologies such as attention techniques, convolution neural networks, and autoencoders for combining disparate medical information modalities in an end-to-end system. We provide a novel taxonomic hierarchy for structuring medical information modalities into three levels starting from structured and unstructured data. Furthermore, we provide a BI tool for diving into MRL-based techniques for medical applications, opening up the potential for researchers to investigate the current state of the art and novel ideas for medical MRL.

The rest of the paper is structured as follows. In Section 2, we introduce preliminary definitions, including the three classification spaces. In Section 3, we introduce the online BI tool. We present a review of previous MRL surveys in Section 4 and conclude in Section 5.

## 2 Classification Spaces

The following classification spaces are used throughout this work to structure our survey of previous work utilizing MRL for ML-based medical analytics. We present three orthogonal dimensions of classification. The *utility* dimension (Section 2.1) describes the intended use of the medical analytics task. The *information modality* dimension (Section 2.2) describes the types of medical information modalities incorporated in the MRL approach. Finally, the *MRL approach* dimension (Section 2.3) describes the MRL technique being used. We begin, however, by defining a medical analytics task.

### Definition E.1 (Medical Analytics Task)

A *medical entity* is an item of interest for medical purposes. Entities can be patients, diseases, tumors, viruses, blood samples, etc. A *medical analytics task* provides information on a medical entity in an automated manner using an algorithm or a learned model over some input data. For example, an ejection-fraction algorithm is designed to calculate the fraction of blood entering the

heart ventricle that is successfully ejected in each cycle using two ultrasound images.

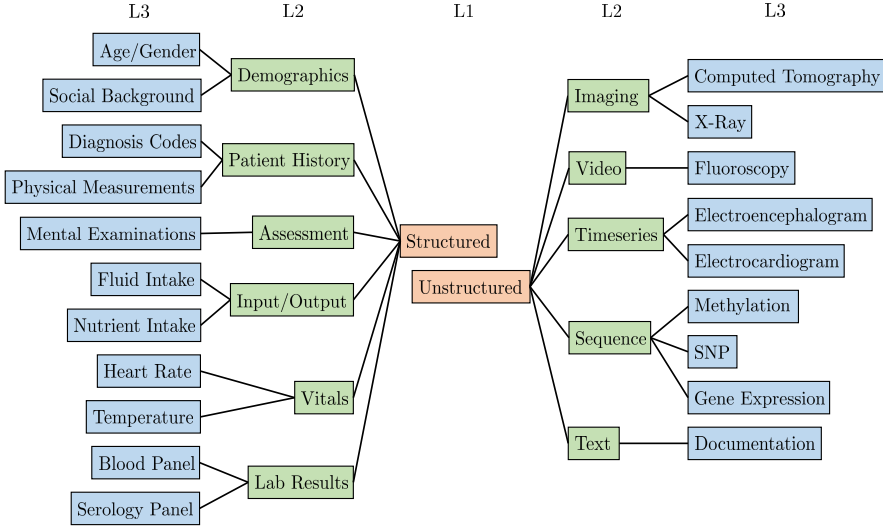
### 2.1 Medical Analytics Utility

It is common to classify *analytics* into one of the following categories of utility [28]. *Descriptive* analytics, as the name suggests, describe the given input. This type of analytics is the most commonly used. It contains methods such as classification (this is an ultrasound image) and object detection (the image contains three lesions in these spatial coordinates). *Diagnostic* analytic methods attempt to identify the root cause of the observed phenomena and are widely used to diagnose diseases and sub-types of diseases with similar symptoms but slightly different causes. *Predictive* analytics, also often described in the medical domain as *prognostic*, attempts to predict the occurrence of a future event or state from the current state or the sequence of states given as input. A medical example could be a sepsis mortality prediction [23]. *Prescriptive* analytics are the most ambitious of the categories, providing one or more recommended actions to take in response to the given input. While in the general domain, this form of analytics is sometimes employed as autonomous agents (e.g., ad recommendation systems), in the medical domain, they are used in a decision support capacity, e.g., treatment recommendation [33].

### 2.2 Medical Information Modalities

A *medical information modality* is a data representation that can be used to obtain information about a medical entity and used by medical analytics. In data management and machine learning [9], it is common to distinguish between structured and unstructured data. Structured information is discretized into records, each containing fields that are assigned values describing some medical entity. For example, a relational database, where tables contain records sharing a fixed schema for describing the status of a patient. However, more flexible data formats such as JSON and XML allow the schema (field names and types) to vary between records.

In the medical domain, we would consider *vitals*, *lab results*, and *demographics* as examples of structured data, as these are mostly single or multi-valued data with a clear interpretation. Unstructured data contains (often large amounts of) data points that hold no inherent meaning when taken individually but that could be interpreted to glean information. For example, the collection of numbers representing an *X-Ray image* makes no sense, but presented as a picture, a trained professional could surmise that it shows a broken tibia. *Images*, *videos*, *timeseries*, *(genomic) sequences*, and *text* are additional examples of unstructured data. While several medical terminologies



**Fig. E.1:** Partial hierarchy for structuring medical modalities. The full hierarchy can be found at <http://tabsoft.co/3DED1Sq>.

exist, e.g., SNOMED [31] and ICD-10 [24], to the best of our knowledge, there exists no hierarchical classification of the medical information modalities that would allow us to perform a structured analysis of the classification. We, therefore, provide a three-level hierarchy of medical information modalities.

Figure E.1 presents a partial view of the proposed hierarchy, showing levels one and two in full and examples from the third level. The top level of the hierarchy separates structured and unstructured modalities. On the second level, we group multiple medical modalities using groupings that are common in ML literature, such as *image*, *text*, and *timeseries*. The third level represents specific medical information modalities used in MRL analytics and can be used to identify which modalities are used by each surveyed analytic approach. Furthermore, to the extent possible<sup>1</sup>, we map our level-three-concepts to SNOMED taxonomy concepts. Hence, the SNOMED subclasses can serve as the fourth and onward levels of the hierarchy if needed. Furthermore, through these concepts, our third-level concepts can be connected to other terminologies using new or established taxonomy mappings to SNOMED and translated to other languages using SNOMED’s language mappings. As an example, our level three concept *Computed Tomography* is mapped to the SNOMED concept *Computed tomography (procedure)*. Using the Biportal SNOMED ontology<sup>2</sup> the SNOMED concept can be mapped to other

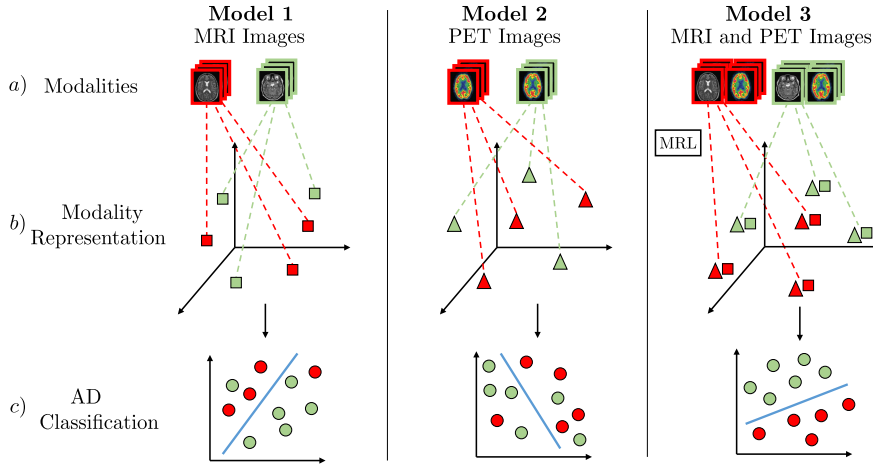
<sup>1</sup>Of the 71 level 3 modalities we identified, 51 were successfully mapped to an equivalent SNOMED concept.

<sup>2</sup><https://biportal.bioontology.org/ontologies/SNOMEDCT>



taxonomies like MedDRA<sup>3</sup> with the concept *CT scan* and BIM<sup>4</sup> with the concept *Computed\_Tomography*. The full hierarchy and its mapping to SNOMED have been made available online at <http://tabsoft.co/3DED1Sq>. Furthermore, the medical modality hierarchy is complete with respect to the set of publications surveyed in this work as described in Section 3.

## 2.3 Multi-modal Representation Learning



**Fig. E.2:** MRL for discriminating AD patients from healthy subjects (HS). The three models (1 and 2 - uni-modal, 3 - multi-modal) consist of three steps: *a)* receive multi-modal samples, *b)* map samples to their individual representation spaces or, in the case of Model 3, use MRL to map modalities to a shared semantic space, *c)* classification of AD/HS. Red represents AD-positive samples, and green represents AD-negative samples. In Models 1 and 2, MRI and PET images are used for uni-modal AD classification. In Model 3, MRL is used to find a shared semantic space combining MRI and PET images, capturing the underlying semantic correlation between these modalities. As illustrated by step *c)* of Model 3, the combined discriminative information from a shared semantic space between MRI and PET can be used for superior medical analytics. See [30] for an example of AD classification using multi-modal MRI and PET images.

To perform a structured analysis of existing MRL approaches, we organize them in a hierarchical structure. Let us first define MRL.

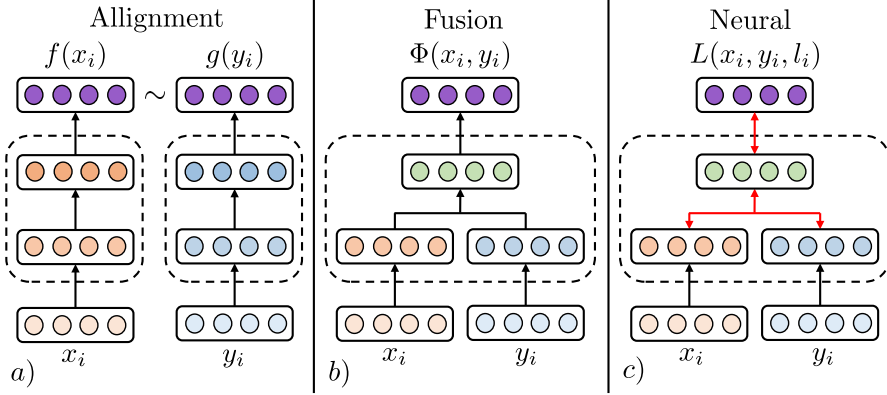
### Definition E.2 (Multi-Modal Representation Learning)

Given two datasets  $x$  and  $y$  of disparate but correlated information modalities, where  $x_i \in x$  and  $y_i \in y$  represent samples describing the same real-world entity, then *multi-modal representation learning (MRL)* is defined as the challenge of finding a latent space where uni-modal modalities can coexist.

<sup>3</sup><https://www.meddra.org/> - MedDRA® trademark is registered by ICH

<sup>4</sup><https://biportal.bioontology.org/ontologies/BIM>

Thus, the latent space contains information from both medical modalities and hence should enable improved subsequent medical analytics compared to uni-modal approaches. The concept is illustrated in Figure E.2, using the example of Alzheimer’s Disease (AD) classification. MRL techniques can



**Fig. E.3:** An illustration of fusion and coordination categories of MRL, adapted from [22]. The MRL step of Model 3 from Figure E.2 can be substituted by techniques from all three categories of MRL.  $x_i$  and  $y_i$  are disparate but correlated uni-modal samples describing the same real-world entity. Arrows represent data transformations, dashed lines are optional transformation steps, and colored dots represent features of  $x_i$  and  $y_i$ . Figure E.3a) illustrates *alignment* MRL, where  $x_i$  and  $y_i$  are aligned through the coordination operator  $\sim$  on  $f(x_i)$  and  $g(y_i)$ . Figure E.3b) illustrates *fusion* MRL, where uni-modal features from  $x_i$  and  $y_i$  are fused through a vector combination technique  $\phi$ . Figure E.3c) illustrates *neural* MRL, where neural network technologies combined with a loss function  $L$  are used to simultaneously learn uni-modal latent representations, a shared latent representation and a medical analytic based on it. Red arrows indicate representation updates using backpropagation.

be broadly classified into *alignment*, *fusion*, and *neural* as illustrated in Figure E.3. Generally, *alignment* techniques find a feature space where modalities can coexist, *fusion* techniques combine uni-modal features into a new latent representation, and *neural* techniques jointly learn a latent representation combining uni-modalities and learn a model for solving a medical analytic. In Section 2.3, Section 2.3, and Section 2.3, we present subcategories of *alignment*-based, *fusion*-based, and *neural*-based MRL techniques respectively.

### Alignment MRL (AMRL)

AMRL learns a representation space in which uni-modal modalities  $x$  and  $y$  can co-exist, with the goal that similar samples should be closer together in the learned space than dissimilar samples. This can be mathematically formulated as  $f(x_i) \sim g(y_i)$ , where  $f$  and  $g$  are modality-specific projection functions that map individual samples  $x_i$  and  $y_i$  into a multi-modal space and

## 2. Classification Spaces

$\sim$  indicates that some distance measure aligns the new space, as illustrated in Figure E.3a. We subdivide AMRL into *correlation* and *similarity* techniques.

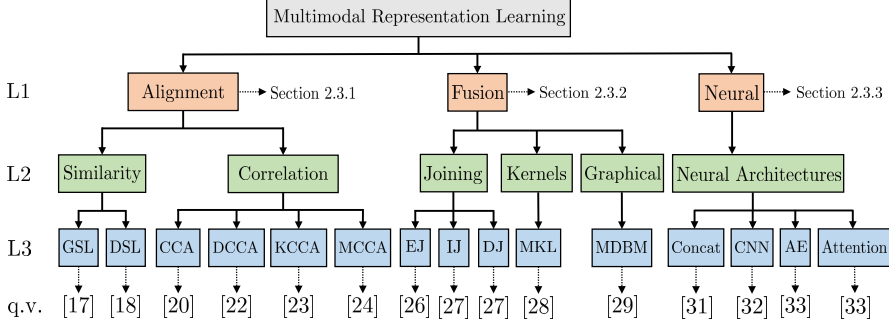


Fig. E.4: Classification space of reviewed MRL techniques used in medical analytics.

**Similarity**-aligned representation learning learns an aligned space between  $x$  and  $y$  by optimizing a distance function for positive and negative modality samples [3]. Shared for all similarity-based methods is the idea of learning transformation matrices  $f$  and  $g$  by minimizing a distance metric such as dot-product similarity or hinge rank loss, often by utilizing stochastic gradient descent (SGD) (Figure E.5). One of the earliest examples is *general similarity learning* (GSL) [35]. GSL creates an aligned space between pairs of images and textual annotations by learning projection functions to map the modalities into a shared space using the weighted approximate-rank pairwise loss. In the resulting coordinated space, similar samples of images and textual annotations will have a smaller cosine distance from each other.

Whereas GSL is limited by choice of initial uni-modal embeddings, *deep similarity learning* (DSL [10]) jointly learns initial uni-modal feature representations and subsequent transformation matrices  $f$  and  $g$  in an end-to-end framework. This can be achieved by adding layers of trainable fully connected neural networks (NN) to step  $a$  in Figure E.5.

Hence, the initial embeddings and subsequent aligned representation space can be jointly learned. Extensions of DSL include using different combinations of loss functions [21] and diverse neural network architectures for uni-modal data transformations.

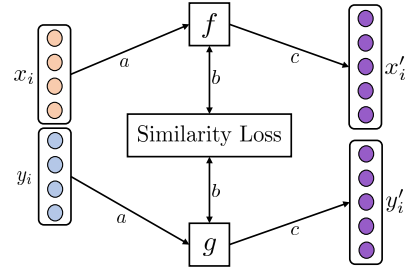
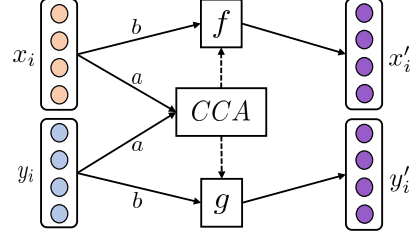


Fig. E.5: GSL technique of AMRL.  $a$  - Modality samples  $x_i$  and  $y_i$  are transformed by modality-specific transformations  $f$  and  $g$ .  $b$  - Using a similarity loss between  $f(x_i)$  and  $g(y_i)$ , SGD iteratively updates  $f$  and  $g$ .  $c$  - When learning has finished,  $f$  and  $g$  transform entities  $x_i$  and  $y_i$  into the coordinated space  $x'_i$  and  $y'_i$ .

**Correlation** entails a set of statistical methods for finding the correlation between two sets of variables. One of the most popular techniques is canonical correlation analysis (CCA). CCA was first introduced in 1936 by H. Hotelling [16]. Given two sets of variables  $x$  and  $y$ , CCA finds the linear projections  $f$  and  $g$  that maximize the correlation between variables from the projected space of  $f(x)$  and  $g(y)$  as  $\arg \max_{f,g} \text{corr}(f(x), g(y))$  as illustrated in



**Fig. E.6:** CCA technique of AMRL. *a* - CCA finds linear transformations  $f$  and  $g$  for uni-modal samples  $x_i$  and  $y_i$  that maximize their projected correlation. *b* - The linear transformations  $f$  and  $g$  are used to project uni-modal samples into the new correlation-optimized aligned space.

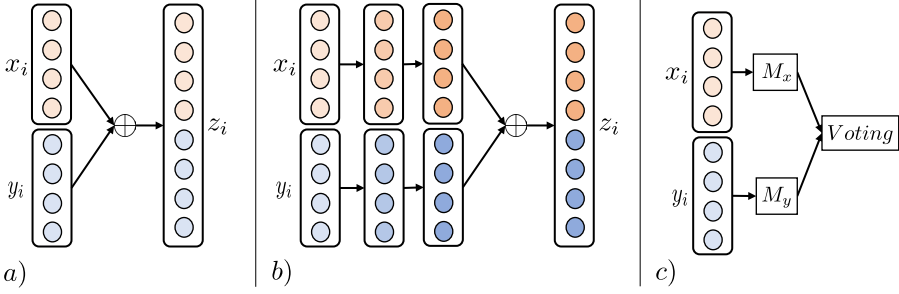
Figure E.6. Finding the transformations resulting in a maximally correlated space can be solved by generalized eigen-decomposition. CCA is thus able to find the linear transformations  $f$  and  $g$ , which maximize the correlation between variables of the transformed modalities. Hence, the original CCA technique is linear with respect to the projection matrices  $f$  and  $g$ , in which case non-linear relationships will not be found. Various extensions to the classical CCA have been proposed to discover non-linear relationships, such as Deep CCA (DCCA) [1] using Fully Connected Neural Networks (FCNNs) for initial feature learning and Kernel CCA (KCCA) [15] utilizing kernels for non-linear feature transformation. Furthermore, extensions to multiple sets of variables have also been proposed, such as Multi CCA (MCCA) [19], which learns a shared space between multiple sets of variables. A review of the relationships between the many CCA variants can be found in [39].

### Fusion MRL (FMRL)

Mathematically, FMRL can be formulated as  $z = \phi(x_i, y_i)$  where  $\phi$  is a function that combines uni-modal data samples  $x_i$  and  $y_i$ , and  $z$  is the combined multi-modal representation. Fusion techniques are usually used to increase the accuracy of classification problems where multiple modalities have distinct discriminative properties [38]. We further divide FMRL techniques into *joining*, *kernels*, and *graphical models*, with complexities varying from linear feature concatenation to complex kernel combinations.

**Joining** combines modalities by concatenating early, intermediate, or late modality-specific features. *Early Joining (EJ)* [2] combines modality features using concatenation functions before any data transformations have been applied to individual modalities as illustrated in Figure E.7a). While EJ is simple and efficient in combining multi-modal data, problems arise when modalities have varying sampling rates. For example, in the combination of

## 2. Classification Spaces



**Fig. E.7:** *a)* Illustrates *EJ* FMRL. Features of uni-modal samples  $x_i$  and  $y_i$  are concatenated through  $\oplus$  to form the fused representation  $z_i$ . *b)* Illustrates *IJ* FMRL, where uni-modal modalities  $x_i$  and  $y_i$  are first processed individually. Later these are concatenated through  $\oplus$ . *c)* Illustrates *DJ* FMRL. Uni-modal modalities  $x_i$  and  $y_i$  are processed through disparate models  $M_x$  and  $M_y$ . Finally, a voting mechanism is applied to the outputs of the individual models.

MRI images and EEG signals. To alleviate such problems, *Intermediate Joining* can be used, where uni-modalities are transformed before latent features are concatenated as illustrated in Figure E.7*b*); however, manual engineering of modality-specific feature transformations is time-consuming and requires extensive domain knowledge.

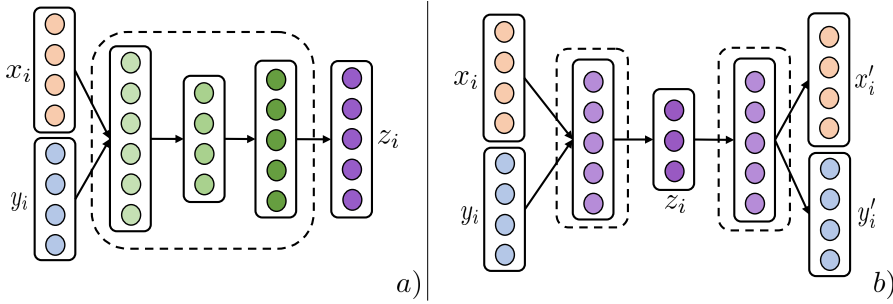
*Decision Joining (DJ)* combines the results of multiple uni-modal analytics, either by majority vote, weighted linear combinations, or more complex techniques as illustrated in Figure E.7*c*). DJ is sometimes preferred in tasks involving low-correlated modalities as the technique is modality independent, and errors from individual analytics tend to be uncorrelated [29].

**Kernels** projects linearly inseparable data into higher dimensional but linearly separable representation spaces using a non-linear kernel transformation. *Multiple Kernel Learning (MKL)* is a sub-type utilizing multiple such kernels. Well-known kernel techniques include support vector machines, the kernel-fisher discriminant, and regularized AdaBoost [12]. Multi-modal representation learning can be achieved by linear, non-linear, or weighted combinations of the resulting modality-specific kernel transformations.

**Graphical Models** are a class of probabilistic machine learning techniques used to discover latent factors explaining the data distribution. Among the most common graphical models for MRL is the *multi-modal deep Boltzmann machine (MDBM)* [32]. An MDBM stacks layers of fully connected restricted Boltzmann machines to form a multi-layer network structure for each modality, which are subsequently joined by an output layer. The idea is to learn a joined density model over the multi-modal inputs such that similarity in the joined space implies similarity between the individual modalities.

### Neural MRL (NMRL)

**Neural Architectures** aim to learn to join representation spaces for multi-modal data in supervised, semi-supervised, or unsupervised ways. Shared for all architectures is the idea of learning layers of non-linear transformations for fusing uni-modal representations into a multi-modal representation space guided by optimizing a loss function [11]. The basis of neural network architectures is the perceptron. The perceptron contains a learnable transformation matrix to linearly transform incoming data modalities into a new representation space, subsequently exercising non-linearity by applying an activation function such as sigmoid or the rectified linear unit.

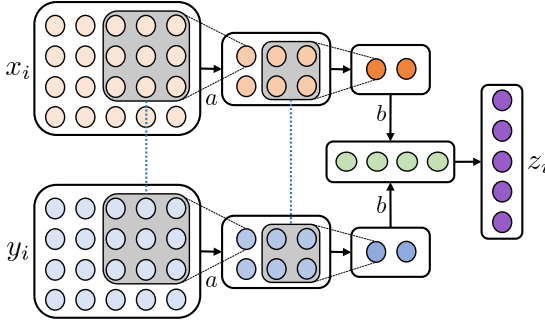


**Fig. E.8:** a) Illustrates the *Concat* technique of NMRL. Uni-modalities are fused by layers of non-linear transformations between input modalities  $x_i$  and  $y_i$  and the fused output space  $z_i$ . Arrows between two neural layers represent the existence of a connection between each neuron/input from a layer to each neuron/output of the next layer. This is true for all neural architectures. b) Illustration of the AE neural network structure. Uni-modal features  $x_i$  and  $y_i$  are transformed through multiple fully connected neural layers. The AE's middle layer learns a low-dimensional fused representation of uni-modalities  $z_i$  by training the neurons using a reconstruction loss between the original modalities  $x_i$  and  $y_i$  and their corresponding reconstructed representations  $x'_i, y'_i$ .

*Concatenation* (Concat) is the most straightforward neural architecture for multi-modal data fusion [14]. Multiple layers of fully connected perceptrons are used to ultimately fuse uni-modal representations in either early, intermediate, or late layers of the network structure (Figure E.8a).

An *Auto Encoder* (AE) is an unsupervised architecture that utilizes a reconstruction loss to learn low-dimensional entity representations that capture most of the original modality information [17]. Multi-modal AE architectures have three stages as illustrated in Figure E.8b). Modality-specific networks that transform uni-modal modalities are initiated and then joined by an intermediary layer that acts as the fused modality representation. The last stage of the architecture splits the intermediary layer into uni-modal networks trained on a reconstruction loss between the final representations  $x'_i$  and  $y'_i$  and the initial representations  $x_i$  and  $y_i$ .

### 3. Analyzing Multi-Modal Research Contributions



**Fig. E.9:** Illustration of CNN technique of NMRL. *a* - A 3 x 3 convolution matrix with shared weights (as indicated by dotted blue lines) slide over the two input modalities  $x_i$  and  $y_i$ . *b* - When sufficiently condensed, features from both modalities can be appended for further processing.

A *Convolutional Neural Network* (CNN) is a technique for learning representations of imaging modalities. Due to the essential domain-specific information images contain, CNNs are often used to learn low-dimensional image representations in end-to-end architectures (Figure E.9). CNNs apply layers of convolution matrices and pooling operations to condense images to their essential discriminative features. Due to their properties, they are often used as an intermediary step of imaging processing with subsequent fully connected layers fusing uni-modal entities [13].

*Transformers* (TF) specialize in learning to represent sequence data. Utilizing a powerful component called self-attention, the model learns the relationships between different parts of the input sequence. This allows the model to attend to specific parts of the input sequence while learning a latent representation for each part of the sequence. This technique can be extended to multi-modal networks by using the learned self-attention weights from one modality in the self-attention mechanism of other modalities. In [25], skin lesion diagnosis is performed using an end-to-end transformer neural network for learning a latent representation between images and clinical features.

### 3 Analyzing Multi-Modal Research Contributions

To understand how medical information modalities have been used together with MRL for medical analytic tasks, we performed a comprehensive, structured survey of publications involving MRL and medical analytics tasks. Using the PubMed search engine<sup>5</sup>, we searched for MRL articles targeting medical analytics tasks while following the PRISMA guidelines [26] for structured

<sup>5</sup><https://pubmed.ncbi.nlm.nih.gov/>

surveys as summarized in Figure E.10. Our search terms were ("*joint*" OR "*fusion*" OR "*coordinated*" OR "*alignment*") AND ("*multimodal*" OR "*multi-view*") AND ("*machine learning*" OR "*deep learning*" OR "*representation learning*") and ("*different modalities*" OR "*multiple modalities*") AND ("*machine learning*" OR "*deep learning*" OR "*representation learning*"). We excluded studies on criteria as summarized in the screening step of the PRISMA guidelines as illustrated in Figure E.10. Eventually, we identified 146 eligible publications.

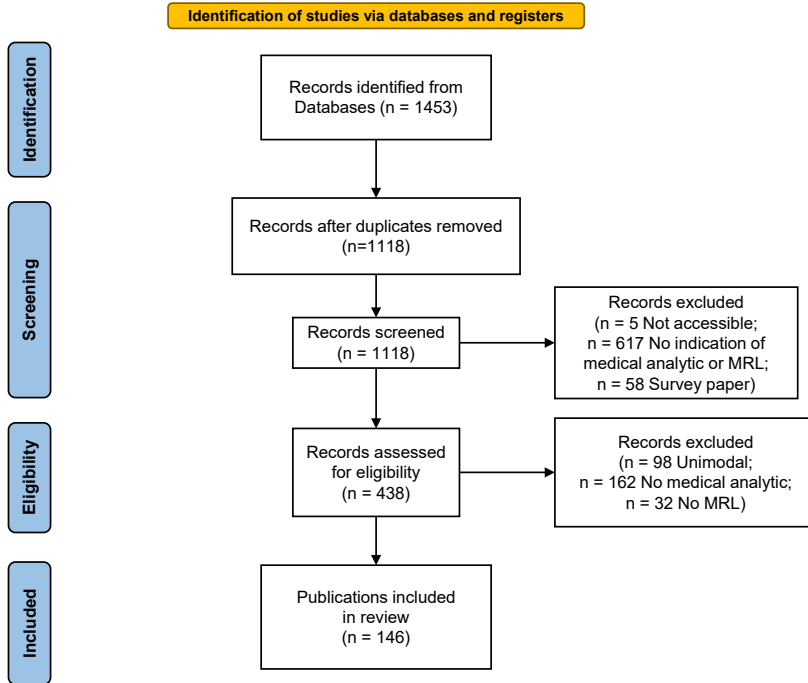


Fig. E.10: PRISMA <sup>6</sup>flow chart for reporting systematic reviews.

To better structure our analysis, we designed four hierarchical dimensions to put the many surveyed papers into a medical and algorithmic context. In total, our analysis employs four dimensions. Namely, *Utility*, *Medical*, *Modality*, and *MRL*. The utility dimension characterizes the analytic utility as described in Section 2.1. The medical dimension uses the ICD-10 [24] diagnosis classification hierarchy to describe the analytic task's medical domain. The modality dimension describes our medical information modality hierarchy as introduced in Section 2.2. The MRL dimension (Figure E.4) describes the MRL technique (Section 2.3). Our primary measure of interest is the number

<sup>6</sup><http://prisma-statement.org>



### 3. Analyzing Multi-Modal Research Contributions

of papers in a specific intersection of dimension values, such as how many papers have used joining MRL for combining structured patient information modalities and unstructured imaging modalities.

Based on the structured survey and our four dimensions of classification, we provide a BI tool together with this work as an electronic supplement <sup>7</sup>. The BI tool can be used to investigate the publications included in this review on our four classification dimensions and provide visual representations of findings. In the remainder of the section, we present a structured analysis of our findings using the BI tool. In Section 3.1, we explore the pairs of modalities observed. In Section 3.2, we examine the prevalence of different MRL techniques in disparate medical fields and for different modality types. In Section 3.3 we examine from the perspective of the different medical analytics tasks encountered.

		Structured						Unstructured			
		Demographics	Input/Output	Lab Results	Structured Assessm..	Structured Patient H..	Vitals	Imaging	Sequence	Timeseries	Video
Structured	Demographics	9	12	8	26			34			
	Input/Output										
	Lab Results	12		9	8	6		21		1	
	Structured Assessm..	8		8		3	1	10			
	Structured Patient ..	26		6	3	10	3	23		5	
	Vitals				1	3				4	
Unstructured	Imaging	34		21	10	23		66		5	
	Sequence										
	Timeseries			1		5	4	5		14	7
	Video									7	

Fig. E.11: Number of papers by level 2 modality combinations used.

#### 3.1 Modality Pairings

Figure E.11 presents the frequency of MRL applications utilizing level 2 (L2) MRL modalities (Figure E.4), combined with level 1 (L1) modalities (Figure E.1). As illustrated, imaging modalities are often combined with other

<sup>7</sup><https://tabsoft.co/3L41do6>

unstructured modalities, specifically other imaging modalities. This is primarily due to medical brain imaging applications, such as AD classification utilizing the distinct discriminative properties of disparate medical imaging technologies. This insight is verified when drilling down into the darkest box (representing imaging-imaging pairings) in Figure E.11 using the Level 3 (L3) modality level. We can see (Figure E.12) that many of these pairs involve PET and MRI scans often utilized in brain studies. A more direct verification can be achieved by adding the medical dimension to this diagram (Figure E.13), where one can see that an overwhelming majority of MRI and PET modalities are used as part of a mental or nervous system analytics task.

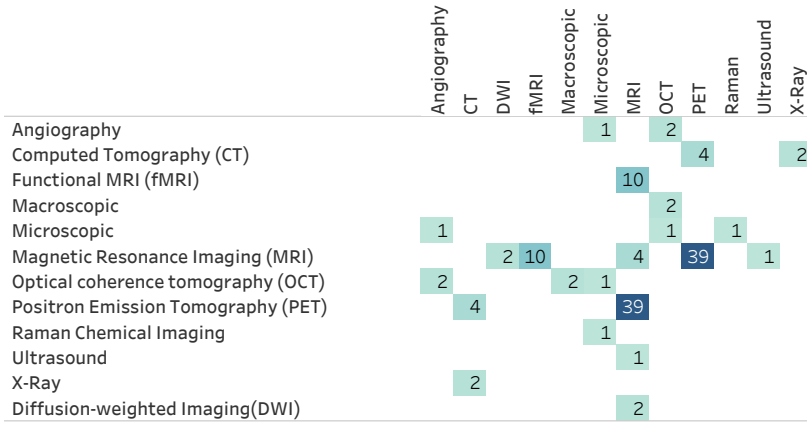


Fig. E.12: Number of papers by level 3 modality combinations, limited to imaging modalities.

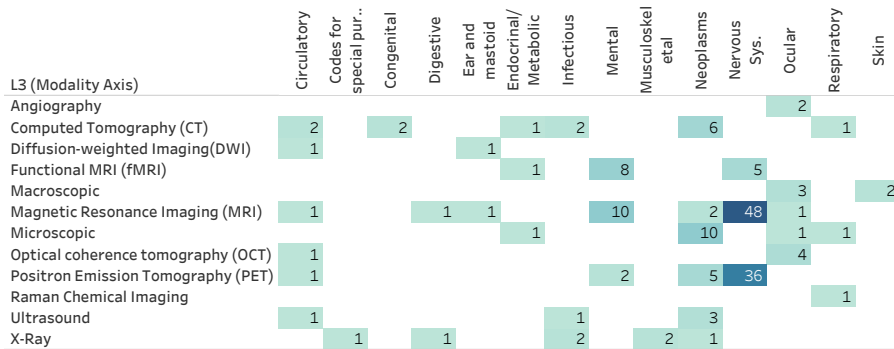
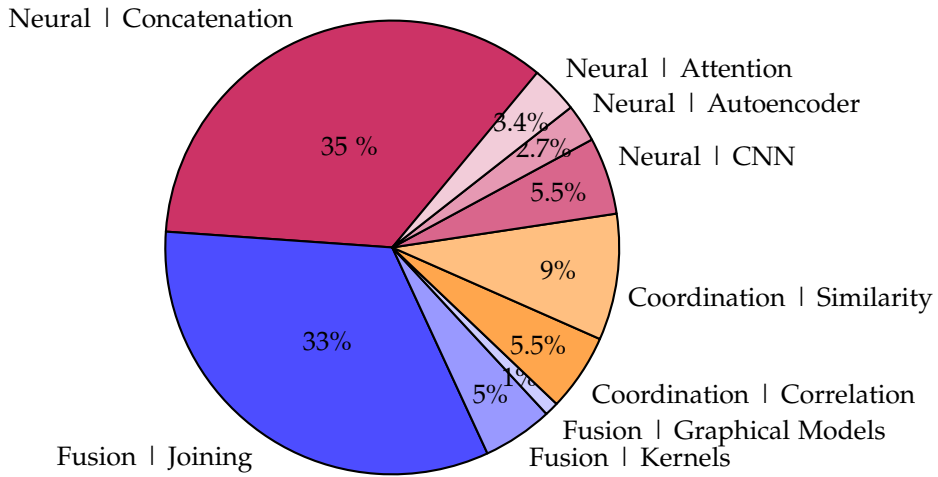


Fig. E.13: Number of papers by level 3 modality and level 1 ICD-10 category, limited to imaging modalities.

Returning to Figure E.11, notice the significantly limited utilization of *structured data* together with imaging and other unstructured modalities such



**Fig. E.14:** Percentage of papers by level-two MRL technique. Techniques in red/orange employ CMRL. The rest of the techniques employ FMRL. Level one followed by level two MRL class is shown for each group of papers.

as audio, graphs, time-series, and video. This could indicate opportunities for future research, as simple structured data, such as demographics, diagnosis codes, and prescriptions, have been shown to increase the discriminative power in multiple medical MRL tasks [7, 18].

## 3.2 MRL Techniques

A hierarchical analysis of the MRL techniques used in literature (Figure E.14) shows that the majority (83,5%) uses fusion MRL techniques, of which the neural architectures and joining L2 types are the most common. Further drill-down into L3 is available in the online supplement BI system.

Comparing the modalities utilized and the MRL techniques employed (Figure E.15), a few results stand out. While neural architectures and joining techniques are evenly used, joining techniques are more prevalent in time-series data and lab results. For time-series, this amounts to 75% of the papers, while in lab results, over 70% of the papers utilize MRL joining techniques. Frequently, medical time series data has an immense sampling frequency leading to hundreds or even thousands of observations per second. Although data transformations can be learned directly for raw time-series data using deep learning techniques such as the artificial recurrent neural network (RNN), the significant sampling rate of modalities like electroencephalography (EEG) can pose algorithmic problems, such as processing time, due to the sheer extent of raw data. This could explain why most time-series data

is processed uni-modally and then fused with other modalities using joining MRL.

L1 (Modality Axis)	L2 (Modality Axis)	MRL L2			
		Graphical Models	Joining	Kernels	Neural Arc hitectures
Structured	Demographics		43%		57%
	Lab Results		71%	6%	24%
	Structured Patient ..		43%		57%
Unstructured	Array		25%		75%
	Imaging	3%	35%	12%	51%
	Timeseries		75%		25%

**Fig. E.15:** Percent of papers utilizing a level 2 MRL technique by level 2 modality. Results are limited to modalities with over 15 papers and to fusion MRL techniques

ICD-10 Cat.	#papers	ICD-10 Cat.	#Papers
Nervous Sys.	55	Musculoskeletal	3
Neoplasms	28	Congenital	2
Mental	20	Skin	2
Circulatory	8	Injury	2
Ocular	6	Ear and mastoid	1
Other	4	Blood diseases	0
Endocrinal/Metabolic	4	Genitourinary	0
Infectious	3	Perinatal	0
Digestive	3	External causes	0
Respiratory	3	Pregnancy and childbirth	0

**Table E.1:** Number of papers by medical task (ICD10 top-level category).

### 3.3 Medical Analytics Tasks

Table E.1 lists the number of publications by medical task (ICD 10 code level 1) in descending order. The names of the categories were shortened for brevity. Thus, *diseases of the eye and adnexa* became *Ocular*. An overwhelming majority of the publications attempt to identify conditions in the nervous system, most commonly the brain itself, as evidenced by 67 of 115 papers being of the nervous system or mental disease categories. Of the remaining categories, neoplasms receive most of the attention, which is an expected re-

#### 4. Related work

sult, given that most of the MRL papers center around imaging modalities. It is somewhat surprising that circulatory system diseases are not commonly addressed as it is a significant focus of medical AI research in general and specifically imaging [5]. Analysis of the type of analytical tasks derived from the MRL revealed that the only two types used were predictive (10) and descriptive (105), as illustrated in Figure E.16.

Machine learning and, in particular, deep learning for medical analytics tasks have recently increased interest. These techniques can learn models directly from labeled data instead of human-engineered feature extraction and modeling techniques. However, the required amount of labeled data needed for training medical analytics in an end-to-end practice exceeds what is readily available for the automated modeling of many medical analytical questions. While some medical analytics tasks have large datasets accessible for immediate consumption in model creation, such as the

Alzheimer’s Disease Neuroimaging Initiative (ADNI) database<sup>8</sup>, researchers are mostly faced with a dearth of annotated datasets [36] significantly limiting the set of problems that can be investigated using ML. MRL still has untapped potential in medical analytics, especially for the less-investigated disease categories of ICD-10, such as dermatological diseases or diseases pertaining to the blood and blood-forming organs. However, ML advances for medical analytics can only progress through considerable data collection processes and open shared access to the collected labeled data repositories.

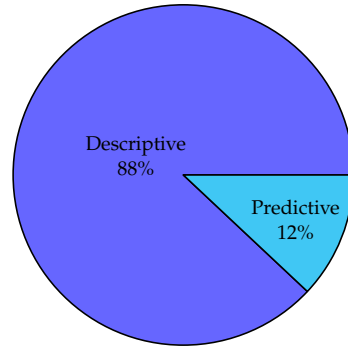


Fig. E.16: Papers by analytic type.

## 4 Related work

Several previous surveys of MRL over general-purpose application domains have focused on the type of MRL technique employed. While surveys reviewing multi-modal deep-learning [14], [34], [38], review new deep learning technologies such as encoder-decoder models, generative adversarial networks, and attention mechanisms, they are narrow in scope of the MRL technique employed and only focus on general-purpose applications. In this work, we do not limit ourselves to specific branches of MRL techniques, but instead do a broad investigation into MRL techniques.

While additional surveys contextualize MRL in the general challenge of multi-modal machine learning [3], with a focus on fusion-based MRL [29],

<sup>8</sup>[www.loni.ucla.edu/ADNI](http://www.loni.ucla.edu/ADNI)

or on its mathematical-theoretical foundations [22], they disregard any specific application domain. We, however, narrow the scope to a comprehensive systematic review of the medical analytics tasks application domain of MRL.

To the best of our knowledge, this paper presents the first attempt to review MRL for medical applications and provide a classification space in which modality combinations from the literature can be placed and future medical analytics designed. Furthermore, we are the first to provide a comprehensive survey on more than 1000 papers for MRL in multi-modal medical applications, while classifying the literature into 4 dimensions of classification e.g., *utility*, *medical*, *modality* and *MRL*.

## 5 Conclusion

In this work, we created a hierarchical taxonomy of medical information modalities linked to the SNOMED concept hierarchy and a hierarchy of related multi-modal representation learning techniques. Subsequently, we performed a literature review of nearly 1100 papers, following the PRISMA guidelines for structured surveys, using MRL to integrate multi-modal medical modalities while inserting these into the four orthogonal dimensions of classification; *utility*, *medical*, *modality*, and *MRL*). Using our classifications, we constructed a free and publically available BI tool for investigating what modalities have been used in combination with each other, for which medical analytics, and what MRL techniques have been successful in what combinations.

We found that a few classes of ICD-10 top-level category disease codes had been the primary target for multi-modal medical analytics. Many ICD-10 classes had only a few to no cases of medical analytics using multi-modal data. We hypothesize that this could be due to the scarcity of openly available labelled training data for medical analytics, forcing ML research to progress in the direction of the medical analytics for which such data is readily available.

While some medical information modalities can be integrated using most MRL techniques, modalities like timeseries need special attention since problems can arise in utilizing end-to-end learning using neural architectures when sampling rates are too high. Before subsequent MRL integration, pre-processing of timeseries modalities using uni-modal techniques is often required.

Furthermore, investigations of the utility dimension show that most medical applications have been developed for descriptive analytics and only a few for predictive analytics. This finding suggests that we are still in an early phase of adopting ML for medical analytics and opens the door for future work in developing prescriptive and even cognitive utility analytics.

As indicated by our systematic review, there is still substantial potential in developing medical analytics using combinations of multi-modal medical information modalities.

## 5.1 Future Work.

To the best of our knowledge, we are the first to do a thorough review of MRL techniques for solving medical ML tasks. In future work, we aim to utilize our four dimensions of classification together with known disease biomarkers for automated modality proposals in medical analytics tasks.

The work presented in this paper is part of our ongoing work to investigate and expand upon techniques for solving medical multi-modal ML tasks. Our long term goal is to utilize multi-modal representation learning for solving medical problems and thereby help clinicians and decision-makers in their work, leading to better patient treatment through personalized medicine.

## References

- [1] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML (3)*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 1247–1255.
- [2] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.
- [3] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [4] K. Blennow, "Cerebrospinal fluid protein biomarkers for alzheimer's disease," *NeuroRx*, vol. 1, no. 2, pp. 213–225, 2004.
- [5] G. Briganti and O. Le Moine, "Artificial intelligence in medicine: today and tomorrow," *Frontiers in medicine*, vol. 7, p. 27, 2020.
- [6] Q. Cai, H. Wang, Z. Li, and X. Liu, "A survey on multimodal data-driven smart healthcare systems: approaches and applications," *IEEE Access*, vol. 7, pp. 133 583–133 599, 2019.
- [7] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.
- [8] R. E. Coleman, "Positron emission tomography diagnosis of alzheimer's disease," *PET Clinics*, vol. 2, no. 1, pp. 25–34, 2007.

## References

- [9] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in health-care," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.
- [10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.
- [11] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
- [12] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.
- [13] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.
- [14] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.
- [15] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [16] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.
- [17] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 202–208.
- [18] M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, T. Doman, A. Ravi, M. Liger, and T. A. Kass-Hout, "Improving hospital mortality prediction with medical named entities and multimodal learning," *CoRR*, vol. abs/1811.12276, 2018.
- [19] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
- [20] R. J. Killiany, T. Gomez-Isla, M. Moss, R. Kikinis, T. Sandor, F. Jolesz, R. Tanzi, K. Jones, B. T. Hyman, and M. S. Albert, "Use of structural magnetic resonance imaging to predict who will get alzheimer's disease," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 47, no. 4, pp. 430–439, 2000.
- [21] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.
- [22] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.



## References

- [23] R. P. Moreno, B. Metnitz, L. Adler, A. Hoechtl, P. Bauer, and P. G. Metnitz, "Sepsis mortality prediction based on predisposition, infection and response," *Intensive care medicine*, vol. 34, no. 3, pp. 496–504, 2008.
- [24] W. H. Organization, *The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 1: Tabular List*. World Health Organization, 2004, vol. 1.
- [25] C. Ou, S. Zhou, R. Yang, W. Jiang, H. He, W. Gan, W. Chen, X. Qin, W. Luo, X. Pi *et al.*, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, 2022.
- [26] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105906, 2021.
- [27] P. S. Pillai and T. Leong, "Fusing heterogeneous data for alzheimer's disease classification," in *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19-23 August 2015*, ser. Studies in Health Technology and Informatics, vol. 216, 2015, pp. 731–735.
- [28] P. Pospieszny, "Software estimation: towards prescriptive analytics," in *Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement, IWSM-Mensura 2017, Gothenburg, Sweden, October 25 - 27, 2017*, M. Staron and W. Meding, Eds., 2017, pp. 221–226.
- [29] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.
- [30] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, and P. Shen, "An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis," *Frontiers in Digital Health*, vol. 3, p. 19, 2021.
- [31] K. A. Spackman, K. E. Campbell, and R. A. Côté, "Snomed rt: a reference terminology for health care." in *Proceedings of the AMIA annual fall symposium*, 1997, p. 640.
- [32] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 2231–2239.
- [33] L. Wang, W. Zhang, X. He, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 2447–2456.

## References

- [34] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 1083–1092.
- [35] J. Weston, S. Bengio, and N. Usunier, "WSABIE: scaling up to large vocabulary image annotation," in *IJCAI*. IJCAI/AAAI, 2011, pp. 2764–2770.
- [36] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed *et al.*, "Do no harm: a roadmap for responsible machine learning for health care," *Nature medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.
- [37] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative *et al.*, "Multimodal classification of alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.
- [38] S. Zhang, J. Zhai, B. Xie, Y. Zhan, and X. Wang, "Multimodal representation learning: Advances, trends and challenges," in *ICMLC*. IEEE, 2019, pp. 1–6.
- [39] X. Zhuang, Z. Yang, and D. Cordes, "A technical review of canonical correlation analysis for neuroscience applications," *Human Brain Mapping*, vol. 41, no. 13, pp. 3807–3833, 2020.



ISSN (online): 2446-1628  
ISBN (online): 978-87-7573-641-6

AALBORG UNIVERSITY PRESS