BHT-Logo-Basis.pdf

Generating Electronic Medical Records

vorgelegt von

Sebastian Herrmann

EDV.Nr.:852049

dem Fachbereich VI
der Beuth Hochschule für Technik Berlin vorgelegte Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science (M.Sc.)
im Studiengang
Medieninformatik

Tag der Abgabe 6. Januar 2020

Gutachter

Prof. Dr.-Ing. habil. Alexander Löser	Beuth Hochschule für Technik Berlin
Prof. Dr. Felix Bießmann	Beuth Hochschule für Technik Berlin

Abstract

XXX

Contents

Chapter 1

Introduction (total approx. 5 pages)

1.1 Problem

1.2 Goal

1.3 Method

1.3.1 Preliminary: GAN

TODO

1.3.2 medGAN

TODO

1.3.3 binary/count variables

For our experiments, we generated both, records with binary and count variables.

1.3.4 privacy risk

TODO

1.4 Overview

Chapter 2

Related Work (approx. 5-8 pages)

2.1 Abstract

The following section will elaborate the basic concepts that are needed to generate synthetic electronic health record data. First, we explain the concept of a Generative Adversarial Network, then we explain medGAN in detail and discuss alternative, superior versions that were released recently.

2.2 Basic Concepts

2.2.1 Generative Adversarial Networks

This section will explain the concept of Generative Adversarial Networks and their role for this work. In **?** proposed a new framework for generative models, that learns the patterns in the data and generates new data that plausibly could originate from the original dataset. The model corresponds to a two-player minimax game in which two independent neural networks train simultaneously: a generator **G** which is learning the distribution of the given data and a discriminator **D** which aims to distinguish between the data from the training set and from **G**.
In this game, G has the goal to maximize the probability of D making a mistake. (**?**)
The two-player minimax game can be described by the following value function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data(x)}}[logD(x)] + \mathbb{E}_{z\sim_{P_z(z)}}[log(1-D(G(z)))] \tag{2.1}$$

To understand the process of the model the following analogy can be helpful: “The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.” (**?**)

2.2.2 Electronic Medical Records

In this section, we will learn the definition of Electronic Medical Records, and how they are distinguishable to Electronic Health Records

2.2.3 Gender Medicine

Medical research is dominated by the male gender, meaning that women are heavily underrepresented or sometimes even excluded from research studies not only in animal studies but also in human trials. **?** But "diseases differ between men and women in terms of prevention, clinical signs, therapeutic approach, prognosis, psychological and social impact."

(**?**) Sex-differences can also be found in the correlation of diseases. (**?**) shows in his research, that "Sex-specific differences appear particularly relevant in the management of type 2 diabetes mellitus (T2DM), with women experiencing greater increases in cardiovascular morbidity and mortality than do men." (**?**) Gender however does not only include sex, but also lifestyle-related diseases, stress and behaviour, like for example regarding help-seeking actions. While we can not take the socio-cultural aspect of gender into account, the differences in sex are applicable. As "cardiovascular disease is the leading cause of death of both men and women" (**?**), we can find numerous occurrences in the MIMIC-III dataset. This allows us to investigate co-occurrences of diabetes and CVD. Originally MedGAN was trained with male and female patients simultaneously, making no difference between them. In this work we are separating the dataset in order to introduce a distinction between genders.
The previously mentioned gender-differences lead us to one of our hypotheses, that will be formulated in Chapter 4.

2.2.4 medGAN

The wide adoption of the electronic health record system by healthcare organizations (HCOs) promises advances in analyzing patient data and computational health. The records however are not easily accessible for researchers. Due to the fact that EHR data consists of personal and sensitive information, access is restricted in order to not induce a privacy risk. Further, to minimize the risk of data misuse, access to such data is regulated by the HCOs. (**?**) (Even researches that are in a direct cooperation with a hospital, do not get access to patient data.) As (**?**) states, "[t]he review process by legal departments and institutional review boards can take months, with no guarantee of access (Hodge Jr et al., 1999). This process limits timely opportunities to use data and may slow advances in biomedical knowledge and patient care (Gostin et al., 2009)." (**?**) describes, that "HCOs often aim to mitigate privacy risks through the practice of de-identification (for Civil Rights, 2013), typically through the perturbation of potentially identifiable attributes (e.g., dates of birth) via generalization, suppression or randomization. (El Emam et al., 2015) However, this approach is not impregnable to attacks, such as linkage via residual information to re-identify the individuals to whom the data corresponds (El Emam et al., 2011b). An alternative approach to de-identification is to generate synthetic data (McLachlan et al., 2016; Buczak et al., 2010; Lombardo and Moniz, 2008). However, realizing this goal in practice has been challenging because the resulting synthetic data are often not sufficiently realistic for machine learning tasks." **?** To overcome the limitations and risks of the above stated methods, (**?**) introduced medGAN, which implements a Generative Adversarial Network that leverages an autoencoder to overcome its limitations: the GAN generates distributed representations of patient records, while the autoencoder decodes them into actual discrete records. This principle and the detailed architecture of medGAN will be further explained in the section 'Electronic Medical Record Generation'.

2.3 Summary

Chapter 3

Data Analysis

Before conducting our research we performed an exploratory data analysis on the MIMICIII dataset. The goal of this analysis was to discover patterns and correlations in the data to formulate hypotheses for further analysis. "MIMIC-III [2] is a freely accessible dataset developed by the MIT lab for computational physiology utilized by numerous researchers and scholars world wide. The database consists of demographics, vital signs, laboratory tests, medications, caregiver notes, and mortality of approximately 40,000 intensive care unit patients at Beth Israel Deaconess Medical center between 2001 and 2012." (**?**) We used pandas and numpy for the analysis and matplotlib, to display the results.
The following section will display the results...
The total number of unique patients found in the dataset is 46260, consisting of 26121 male of 20399 female patients.
The correlation heatmap that is introduced in (**?**) shows the strongest correlation of ICD9 codes from the groups two and six which correspond to endocrine, nutritional and metabolic diseases and immunity disorders and diseases of the circulatory system.
We further investigated these groups and focus on comparing the cooccurrence of ischemic and diabetic diseases.
First, we compare the top 12 occcurences of male and female diagnoses of heart disease
Out of 46260 patients, in our dataset 14851 female and 19185 male are diagnosed with any form of heart disease. We can see, that mostly women are affected by Takotsubo syndrome (ICD9-Code 429.9) and that around twice as many men are affected by Rupture of chordae tendineae as women (ICD9-Code 429.5). For unspecified Heart diseases (code 429.9) and other ill-defined heart diseases (code 429.89) more men are affected. However we can find more women affected by functional disturbances following cardiac surgery and unspecified Myocarditis.
Second, we compare males and females diagnosed with diabetic diseases: (ICD9 250.*). 7280 females and 9174 males are affected.
As explained in [Basic concepts - Gender medicine] correlations between diabetes and heart disease find their cause in this subject

Table 4. Top frequent ICD codes of MIMIC-III

Top ICD codes	Meaning	Frequency	No. of patients associated with	Percent of patients associated with
ICD_401	Essential hypertension	21 329	18 031	38.76 %
ICD_427	Cardiac dysrhythmias	20 998	14 022	30.14 %
ICD_428	Heart failure	20 676	10 154	21.83 %
ICD_276	Disorders of fluid, electrolyte, and acid-base balance	20 440	12 645	27.18 %
ICD_250	Diabetes mellitus	16 454	10 318	22.18 %
ICD_414	Other forms of chronic ischemic heart disease	15 759	11 926	25.64 %
ICD_272	Disorders of lipoid metabolism	14 768	12 268	26.37 %
ICD_518	Other diseases of lung	14 608	11 363	24.43 %
ICD_285	Other and unspecified anemias	12 910	10 631	22.85 %
ICD_584	Acute renal failure	11 467	9536	20.50 %

Figure 3.1: The traditional supervised learning setup in machine learning. (**?**)

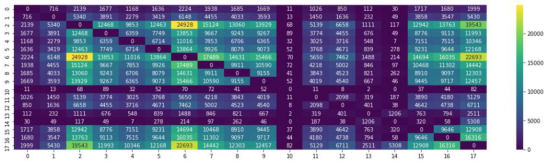

Figure 3.2: The traditional supervised learning setup in machine learning. (**?**)

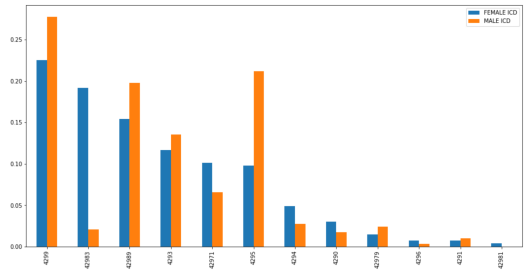

Figure 3.3: Comparison male/female heart disease. The X axis describes the ICD9 code, the Y axis the percentage of affected patients.

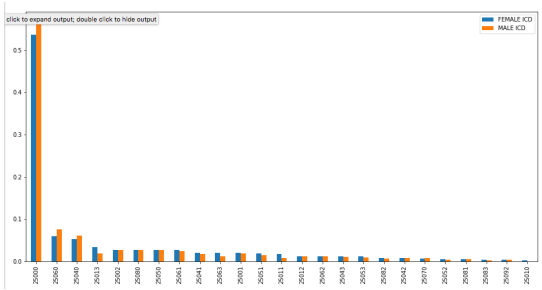

Figure 3.4: Comparison male/female diabetic disease. The X axis describes the ICD9 code, the Y axis the percentage of affected patients.

Chapter 4

Electronic Medical Record Generation (approx. 15 pages

4.1 Abstract

In this chapter, we will elaborate medGAN, discuss the process and architecture, our experimental setup and the process of training and generating synthetic Electronic Medical Record data. Further, we will introduce our hypotheses and describe the dataset.

4.2 medGAN

D and G are both implemented as feedforward neural networks. As we learned in (SECTION GAN), the generator G "is trained by the error signal from the discriminator D via backpropagation, the original GAN can only learn to approximate discrete patient records x Z|C| with continuous values. " (**?**)
To alleviate this limitation, they leveraged an autoencoder which is reconstructing an dimensionality reduced approximate of the input. As (**?**) stated, "[s]uch a mechanism leads the autoencoder to learn salient features of the samples and has been successfully used in certain applications, such as image processing (Goodfellow et al., 2016; Vincent et al., 2008)."
The objective of the autoencoder is, to minimize the reconstruction error:
FORMEL AUTOENCODE

$$\frac{1}{m}[\sum_{i=0}^{m} || \ x_i - x_i` \ ||_2^2] \tag{4.1}$$

$$\frac{1}{m}[\sum_{i=0}^{m} x_i \log x_i + (1 - x_i) \log(1 - x_i)] \tag{4.2}$$

FORMELN STIMMEN NOCH NICHT where m is the size of the mini-batch.
An autoenconder consists of of two elements: The Encoder (Enc) which compresses the input and the Decoder (Dec) that is used to construct the output. For count variables they used the cross entropy loss and rectified linear units as activation function for the Enc and Dec. For binary variables they used the mean squared loss and the tanh activation function for Enc and the sigmoid activation for Dec. Both, the Enc and the Dec are implemented as single layer feedforward networks. The original input x it receives, is compressed into a 128 dimensional vector. The generator G consists of two hidden-layers with each 128 dimensions and is implemented as a feedforward network. For the batch normalization in G they use the scale parameter $\gamma and the shift parameter \beta and set the moving average decay to 0.99. The discriminator D has the same structure, layer consists of 256 dimensions. medGAN is trained for 1,000 epochs with the mini - batch of 1000 records.$(?)

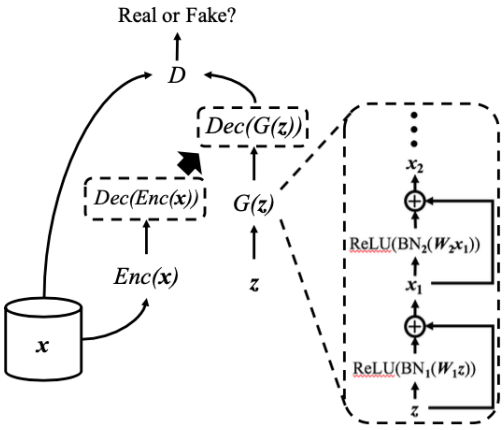

Figure 4.1: Architecture of medGAN: The discrete x comes from the source EHR data, z is the random prior for the generator G; G is a feedforward network with shortcut connections (right-hand side figure); An autoencoder (i.e, the encoder Enc and decoder Dec) is learned from x; The same decoder Dec is used after the generator G to construct the discrete output. The discriminator D tries to differentiate real input x and discrete synthetic output Dec(G(z)). (**?**)

"medGAN generates synthetic EHR datasets that achieve comparable performance to real data on many experiments including distribution statistics, predictive modeling tasks and medical expert review."
UMFORMULIEREN (zitiert)

4.2.1 SynthEHR (medBGAN, medWGAN)

In (**?**) proposed two altered versions of medGAN that outperform their predecessor, however just slightly. Those two versions are:

medWGAN: This version substitutes the regular GAN with an improved *Wasserstein GAN (WGAN)*, that utilizes "an alternative method of weight clipping called gradient penalty, which entails penalizing the norm of the gradient of the discriminator (critic) with respect to its input" (**?**)

medBGAN: This version substitutes the regular GAN as well, but this time with a *boundary-seeking GAN (BGAN)*. This approach trains the generator to match the target distribution that converges toward the true distribution as the discriminator is optimized" (**?**)

4.2.2 Differentiation

Bereits erklärt?

4.2.3 Why left out?

If our hypotheses proof to be true and bring an improvement to medGAN, these improvements will also translate to altered versions of medGAN, because not the network itself is being changed but the input. In our tests we separated the dataset and did not alter it. Henceforth we performed our tests only on the 'original' medGAN.

4.3 Process and Architecture

For generating the Electronic Medical Records (EMR), we used a Generative Adversarial Network (GAN) called medGAN, that was proposed in (**?**). As input data we use v1.4 of the MIMIC-III (Medical Information Mart for Intensive Care) dataset. For our experiments, we divided the dataset by gender and generated EMR data with binary and count variables for both mixed and separated patients. The code for medGAN is publicly accessible under https://github.com/mp2893/medgan. It is implemented using TensorFlow For training models, they chose the Adam-optimizer with a mini-batch size of 100 patients. (**?**) We trained the model using Colaboratory by Google, which is a Jupyter notebook environment, providing free Cloud computing for education and research. The machines are equipped with K80 GPUs from NVIDIA. With the K80 GPU, it took 29 minutes and 3 seconds to train the model with only female patients, 35 minutes and 26 seconds to train the model with only male patients and 60 minutes and 1 second to train the model with the full dataset.

4.4 Experimental Setup

For training, we split the data into subsets with a ratio of 9:1 for training and validation subsets. Using the training subset, the autoencoder is pretrained for 100 epochs. After each epoch, we report the training and validation loss. For binary variables we use the cross-entropy loss function, for count variables the mean squared error. Further, we use minibatch averaging and batch normalization We conducted our data analysis in a Jupyter Notebook. Here, we use pandas to investigate the data and matplotlib to show our results. After finishing the training process, we select the epoch closest to 0.5, since that is when the discriminator is most confused and the generator makes the most convincing synthetic samples.

4.5 Hypothesis

In this work we are trying to proof the following three hypotheses:
First, the model can generate realistic patients if it is trained with the MIMIC III dataset / the model learns the distribution of ICD9 codes. Second, by training the network with female and male patients separately, it is able to generate patients with gender-specific correlated diseases that seem realistic to a medical doctor. Third, if the network is being trained with the MIMIC-III dataset, it is able to generate patients affected by orphan diseases that can not be distinguished to a non-synthetic patient by a medical doctor

4.6 Data

The dataset is publicly available for researchers worldwide. In order to gain access to it we were required to complete the CITI "Data or Specimens Only Research" course.
It contains deidentified health-related data associated of over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012.
We extracted the ICD9 codes for female and male patients from the dataset separately and aggregated a patient's longitudinal record into a single fixed-size vector x elem $Z^C where Cequals 906 formixedpatients, 857 forfemalepatients and 848 formalepatients.$
TODO: GLEICHUNGEN
Model for comparison:
To assess the effectiveness of my method, I trained the network both, with mixed and gender-separated patients.

ICD9-Code	Percent of affected patients
296.20 Major depressive disorder single episode	36.12%
V45.82 Percutaneous transluminal coronary angioplasty status	30.89%
997.39 Other respiratory complications	26.91%
401.9 Unspecified essential hypertension	26.20%
715.90 Osteoarthrosis and allied disorders	26.18%
596.8 Other specified disorders of bladder	23.91%
410.71 Subendocardial infarction initial episode of care	21.02%
774.2 Neonatal jaundice associated with preterm delivery	20.55%
250.11 Diabetes mellitus with ketoacidosis type i not stated as uncontrolled	20.50%
593.9 Unspecified disorder of kidney and ureter	19.90%

Table 4.1: Top 10 diagnoses for generated samples (both sexes).

? also compared medGAN with other popular generative methods like Random Noise, Independent Sampling. In his experiments, medGAN outperformed other methods. Therefore we will compare our results with medGAN.

4.7 Models for Comparison

medGAN ORIGINAL RESULTS

4.8 Measurements

For comparison we compared sample sizes, equal to the dataset.

4.9 Summary

4.10 Non-Gender-specific EMR Generation

First, like in (**?**) we generated EMR data, without dividing the dataset. The results serve as baseline for the performance and will be compared with the separately generated patients.

4.11 Gender-specific EMR Generation

Second, we divided our dataset by gender and generated patient data separately, resulting in 20399 female and 26121 male unique records.

4.12 Orphan diseases

Wikidata provides a mapping of 4584 ICD-9 codes to GARD and OrphaNet IDs. To investigate the occurrence of orphan diseases in MIMIC-III, we first generated a list containing all 6986 unique

ICD9-Code	Percent of affected patients	
4019	20703	
4280	13111	
42731	12891	
41401	12429	
5849	9119	
25000	9058	
2724	8690	
51881	7497	
5990	6555	
53081	6326	

Table 4.2: Top 10 diagnoses MIMIC-III.

ICD9 codes of the dataset. Then, we match the list from the mapping which contains a total of 962 codes, resulting in ten corresponding codes in the MIMIC-III dataset. 2408 diagnoses with orphan ICD9-codes are present in the given dataset.
The following section will elaborate our findings regarding orphan diseases in our generated patients.

Chapter 5

Experimental Evaluation

5.1 Abstract

In the following section we will, evaluate our expermiments in order to measure our success. First, in a quantitative manner, using three statistical methods. Second, by evaluation in a qualitative manner, comparing the generated data to the original dataset and with the help of a medical doctor.

5.2 Hypotheses

In our work we are trying to proof the following hypotheses: First, we are making the assumption that the model can generate realistic patients if it is trained with the MIMIC III dataset / the model learns the distribution of ICD9 codes. Second, by training the network with female and male patients separately, it is able to generate patients with gender-specific correlated diseases that seem realistic to a medical doctor -> improvement - if the network is being trained with the MIMIC-III dataset, it is able to generate patients affected by orphan diseases that can not be distinguished to a non-synthetic patient by a medical doctor

5.3 Measurements

Orphan diseases:

ICD9-Code	Female Binary	Male Binary	Mixed
042 Human immunodeficiency virus [HIV] disease	739	144	121
515 Postinflammatory pulmonary fibrosis	0	168	907
570 Acute and subacute necrosis of liver	255	0	33

ICD9-Code	Dataset Female	Dataset Male
075 Infectious mononucleosis	7	4
138 Late effects of acute poliomyelitis	36	37
193 Malignant neoplasm of thyroid gland	21	28
220 Benign neoplasm of ovary	25	0
317 Mild intellectual disabilities	43	39
8832 Open wound of finger(s), with tendon involvement	2	15

5.4 Evaluation

5.4.1 Abstract

5.4.2 Quantitative Evaluation UMFORMULIEREN

For the quantitative evaluation of our measurements we choose the following statistical methods, as presented in (**?**).

Dimension-wise probability: This refers to the Bernoulli success probability of each dimension (disease or procedure code) in the binary dataset. The dimension-wise probability is computed using the following formula:

$$Number\,of\,patients = \frac{Number\,of\,patients\,who\,had\,the\,disease}{Total\,number\,of\,patients} \tag{5.1}$$

We calculate it for the binary data.

Dimension-wise average: This refers to the column average of each dimension (disease or procedure code) in the count dataset. The dimension-wise average is calculated using the following formula:

$$Dimension - wise\,average = \frac{Column\,sum}{Total\,number\,of\,records} \tag{5.2}$$

We calculate it for the count data.

Dimension-wise K–S test: We performed the K–S test on 2 data samples (synthetic data and real data) to examine whether the 2 data samples originate from the same distribution. In the K–S test, the statistic is calculated by finding the maximum absolute value of the differences between 2 samples' cumulative distribution functions. The null hypothesis is that both samples originate from a population with the same distribution. In our experiment, we rejected the null hypothesis with a low P-value (typically 0.05). More details of the K–S test is discussed in the Results section.

Evaluation Orphan diseases

5.4.3 Qualitative Evaluation

We performed the qualitative evaluation of our measurements with the help of a medical doctor. Therefore we take 50 samples from the original dataset and 50 samples from our generated patients. Then, we present them in random order to a medical doctor and let him rate them on realisticness on a scale from 1 to 10, where 10 is the highest score and therefore a sample, that can not be distinguished from a real record. The samples are selected randomly, except for the orphan diseases. Here we choose 5 samples from the dataset and 5 from our generated patients because of their scarcity.

Gender-specific

First, we compare the top 10 occurring ICD9 codes of our generated samples with those of the original dataset.
Further, we put the focus of our examination on diabetic and ischemic diseases. From our generated samples, 823 females are affected by both types, while 428 males are affected by both. When generating patients, without separation 1646 are affected.
- Men have a higher risk for Coronary atherosclerosis of native coronary artery (ICD 41401) - only women affected: - Chronic diastolic heart failure (ICD9 42832) - Acute on chronic diastolic heart

ICD9-Code	Percent of affected patients
397.0 Diseases of other endocardial structures	32.09%
426.11 First degree atrioventricular block	31.59%
296.20 Major depressive disorder single episode	31.21%
599.0 Other disorders of urethra and urinary tract	27.16%
852.21 Subdural hemorrhage following injury without open intracranial wound with no loss of consciousness	23.24%
V29.0 Observation for suspected infectious condition	20.85%
414.01 Coronary atherosclerosis of native coronary artery	18.94%
427.1 Paroxysmal ventricular tachycardia	18.32%
997.1 Cardiac complications not elsewhere classified	17.96%
770.6 Transitory tachypnea of newborn	17.05%

Table 5.1: Top 10 diagnoses for generated female samples.

ICD9-Code	Percent of affected patients
401.9 Unspecified essential hypertension	40.72%
272.4 Other and unspecified hyperlipidemia	30.82%
414.01 Coronary atherosclerosis of native coronary artery	28.84%
998.11 Hemorrhage complicating a procedure	27.75%
553.21 Incisional hernia without obstruction or gangrene	24.59%
571.5 Cirrhosis of liver without alcohol	24.55%
427.1 Paroxysmal ventricular tachycardia	20.51%
E879.0 Other procedures without mention of misadventure at the time of procedure as the cause of abnormal reaction of patient or of later complication	19.27%
996.04 Mechanical complication of automatic implantable cardiac defibrillator	19.22%
482.41 Methicillin susceptible pneumonia due to staphylococcus aureus	16.96%

Table 5.2: Top 10 diagnoses for generated male samples.

failure (ICD9 42833) - Intracerebral hemorrhage (ICD9 431) - Takotsubo syndrome (ICD9 42983) affects mostly women - rupture of chordae tendineae (ICD9 4295) affects mostly men (connecting mitral valve and tricuspid valve with papillary muscles); -> predominantly among men older than 50 years, and is rare in young adults and children (Source (((1))))

5.4.4 Discussion

5.4.5 Summary

Chapter 6

Conclusion and Outlook (5 pages)

6.1 Goal

Our goal was to proof, that medGAN can generate realistic patients if it is trained with the MIMIC III dataset / the model learns the distribution of ICD9 codes Further we tried to elaborate if by training the network with female and male patients separately, it is able to generate patients with gender-specific correlated diseases that seem realistic to a medical doctor. Also we wanted to find out whether the model is able to generate patients affected by orphan diseases that can not be distinguished to a non-synthetic patient by a medical doctor despite the rare occurrences in the dataset.

6.2 Hypothesis proofed?

6.3 Outlook

6.4 Future work

6.5 Summary

Acknowledgements

I would like to thank Prof. Dr. Alexander Löser for the insightful discussions and valuable comments as well as his support and supervision of this thesis. I also would would like to express my gratitude to my advisor Betty van Aken for her continuous guidance and supervision. Without their help, this thesis would not have been possible.

Appendix A

Datasets

xxx

Appendix B

Listings

xxx

Appendix C

Results

xxx