Universität Bielefeld

Fakultät für Wirtschaftswissenschaften

Masterarbeit

im Studiengang Statistische Wissenschaften

zum Thema:

Classifying Emergency Department Data to Improve Syndromic Surveillance: From Mixed Data Types to ICD Codes and Syndromes

vorgelegt von

Birte Wagner

Matrikel-Nr: 2964907

1. Prüferin: Prof. Dr. Christiane Fuchs
2. Prüfer: Prof. Dr. Roland Langrock

Bielefeld, 6. April 2020

Abstract

Syndromic surveillance systems are used to monitor public health and enable a timely outbreak detection. Emergency department (ED) data can serve as an important data source for syndromic surveillance, but a high amount of missing diagnosis codes can make analyses relying on this information impossible. This study aims at enhancing an ED dataset from a piloted syndromic surveillance system in Germany to enable the monitoring of an influenza-like illness (ILI) syndrome.

Routinely collected data from one ED containing mixed-type variables are analysed and two different approaches are implemented to deal with the missing data. Within the first approach, the missing diagnosis codes are imputed by predicting them from the remaining variables, using a multi-class naive Bayes classifier and a deep learning imputation package. In the second approach, a logistic regression model and a binary naive Bayes classifier are used to predict the ILI syndrome from all variables except the diagnosis code. The resulting ILI cases are evaluated on time series level with regard to seasonal patterns.

The diagnosis codes were predicted from mixed-type input variables with sufficient precision (34.37% F1-measure in the best model). By taking into account the hierarchical structure of the ICD-10 codes, the performance was improved. Predicting the ILI syndrome independent of the diagnosis code from the remaining variables worked well (39.63% F1-measure in the best model) and the predictions showed medical similarity with the ILI syndrome. The models differed in their sensitivity of including cases, which can be adjusted by changing the threshold of the classifiers. The resulting ILI cases from all models were positively correlated with the reference cases on a time series basis ($r = 0.865$ for best model) and were comparable with an external data source, a surveillance of severe acute respiratory infections (SARI) ($r = 0.867$ for best model).

The present study showed that the ED dataset can be enhanced to enable the syndromic surveillance of an ILI syndrome based on the diagnosis codes, even if this variable is missing. Additionally, a flexible case definition for an ILI syndrome was developed that is independent of the diagnosis code and the underlying generic method can be applied to other syndromes as well.

Zusammenfassung

Syndromische Surveillance Systeme werden verwendet, um die öffentliche Gesundheit zu überwachen und ermöglichen eine frühzeitige Ausbruchserkennung. Notaufnahmedaten können eine wichtige Datenquelle für syndromische Surveillance bieten, jedoch kann eine hohe Zahl an fehlenden Diagnosecodes Analysen unmöglich machen, die auf diese Information angewiesen sind. Diese Studie zielt darauf ab, einen Notaufnahme-Datensatz aus einem pilotierten syndromischen Surveillance System für Deutschland zu verbessern, um die Überwachung eines Influenza-like Illness (ILI) Syndroms zu ermöglichen.

Routinemäßig gesammelte Daten aus einer Notaufnahme, die Variablen unterschiedlichen Datentyps enthalten, werden analysiert und zwei verschiedene Ansätze zum Umgang mit den fehlenden Daten implementiert. In dem ersten Ansatz werden die fehlenden Diagnosecodes imputiert, indem sie aus den übrigen Variablen vorhergesagt werden. Dafür wird ein Multiclass naive Bayes Modell und ein Deep Learning Imputations Package verwendet. In dem zweiten Ansatz werden ein logistisches Regressionsmodell und ein binäres naive Bayes Modell verwendet, um das ILI Syndrom aus allen Variablen außer dem Diagnosecode vorherzusagen. Die resultierenden ILI Fälle werden auf Zeitreihenebene in Bezug auf saisonale Muster evaluiert.

Die Diagnosecodes konnten aus den verschiedenen Variablen mit ausreichender Präzision vorhergesagt werden (34.37% F1-Wert im besten Modell). Durch Berücksichtigung der hierarchischen Struktur der ICD-10 Codes konnte das Ergebnis verbessert werden. Die Vorhersage des ILI Syndroms unabhängig von dem Diagnosecode aus den restlichen Variablen funktionierte gut (39.63% F1-Wert im besten Modell) und die vorhergesagten Fälle wiesen eine medizinische Ähnlichkeit zu dem ILI Syndrom auf. Die Modelle unterschieden sich in ihrer Sensitivität, mit der ILI Fälle eingeschlossen werden. Diese kann angepasst werden, indem der Schwellenwert der Klassifikationsmodelle geändert wird. Die resultierenden ILI Fälle aller Modelle waren positiv mit den Referenzfällen auf Zeitreihenebene korreliert ($r = 0.865$ für das beste Modell) und ähnelten einer externen Datenquelle, einer Surveillance von schweren akuten respiratorischen Infektionen (SARI) ($r = 0.867$ für das beste Modell).

Mit dieser Studie wurde gezeigt, dass der Notaufnahme-Datensatz verbessert werden konnte, sodass eine syndromische Surveillance von einem ILI Syndrom auf Basis von Diagnosecodes möglich ist, selbst wenn diese Variable fehlt. Zusätzlich wurde eine flexible Falldefinition für ein ILI Syndrom entwickelt, die unabhängig von dem Diagnosecode ist und dessen zugrundeliegende generische Methode auch auf andere Syndrome angewendet werden kann.

Contents

1 Introduction and motivation **1**
- 1.1 Syndromic surveillance . . . 1
- 1.2 Influenza and influenza-like illness . . . 3
- 1.3 Automatic prediction of clinical diagnoses . . . 4
- 1.4 Syndrome prediction and unsupervised learning of case definitions . . . 6
- 1.5 Goals of this thesis . . . 7

2 Description of the dataset **9**
- 2.1 Data source . . . 9
- 2.2 Variables and preprocessing . . . 9
- 2.3 Exploratory data analysis . . . 10
- 2.4 Expert case definition of the ILI syndrome . . . 14
- 2.5 Comparison with external data sources . . . 15

3 Methods and implementation **17**
- 3.1 Imputation of missing data . . . 17
  - 3.1.1 Introduction to common imputation methods . . . 17
  - 3.1.2 Classification approach . . . 18
  - 3.1.3 DataWig approach . . . 18
- 3.2 Binary and multi-class classification . . . 19
  - 3.2.1 The classification setting . . . 19
  - 3.2.2 Naive Bayes . . . 20
  - 3.2.3 Logistic regression . . . 20
  - 3.2.4 Imbalancedness . . . 21
  - 3.2.5 Evaluation measures . . . 21
- 3.3 Implementation and procedure of data analysis . . . 22
  - 3.3.1 Train- and test datasets . . . 24
  - 3.3.2 Imputation of missing diagnosis codes . . . 24
  - 3.3.3 Classification of an ILI syndrome . . . 25
  - 3.3.4 Evaluation of weekly aggregated cases . . . 27

4 Results and evaluation **29**
- 4.1 Imputation of missing diagnosis codes . . . 29
  - 4.1.1 Evaluation of imputed diagnosis codes . . . 29
  - 4.1.2 Evaluation of the syndrome cases based on imputed diagnosis codes . . . 29
- 4.2 Classification of an ILI syndrome . . . 30
  - 4.2.1 Evaluation of the syndrome predictions . . . 30
  - 4.2.2 Variable importance of the logistic regression models . . . 31
- 4.3 Evaluation of weekly aggregated cases . . . 33
  - 4.3.1 Imputed diagnosis codes . . . 34
  - 4.3.2 Syndrome models . . . 34
  - 4.3.3 Selection of the best models . . . 35
  - 4.3.4 Further evaluation of the selected models . . . 36

5 Discussion **44**
- 5.1 Critical discussion of the results . . . 44
- 5.2 Limitations of this work . . . 48
- 5.3 Conclusion . . . 49

Appendices **54**

A Supporting tables and figures **54**

B Eidesstattliche Erklärung **59**

This thesis was written in the ESEG and Signale project groups at the Robert Koch Institute (Department 3 – Infectious Disease Epidemiology, Unit 31 – Infectious Disease Data Science and Unit 32 – Surveillance).

List of Tables

1.1 ICD-10 chapters . . . 5
2.1 Examples for MTS complaint group and value . . . 9
3.1 Dataset sizes of the original dataset and train- and testsets . . . 24
4.1 Evaluation metrics for diagnosis models . . . 29
4.2 Evaluation metrics for diagnosis models on syndrome level . . . 30
4.3 Evaluation metrics for syndrome models . . . 31
4.4 Correlation of ILI cases with best models and ICOSARI . . . 38
4.5 False positives of best models, by category . . . 41
4.6 False positives of best models, by diagnosis (letter) . . . 42
4.7 False negatives of best models, by category . . . 42
4.8 False negatives of best models, by diagnosis (block) . . . 43
A.1 Variables originally included in the dataset . . . 54
A.2 Descriptive statistics compared for different datasets . . . 55
A.3 Abbreviations in the variable referral . . . 55
A.4 R packages used in this work . . . 58

List of Figures

2.1 Absolute number of missing diagnosis codes vs. cases per week 11
2.2 Percentage of missing values in each variable 11
2.3 Distribution of all metric variables 12
2.4 Frequencies for all categorical variables 13
2.5 Comparison of time series from different external data sources 16
3.1 Confusion matrix 21
3.2 Procedure flow chart 23
4.1 Variable importance of logistic regression models 32
4.2 Weekly aggregated cases of imputation models 33
4.3 Weekly aggregated cases of syndrome classifiers 35
4.4 Weekly aggregated cases of best models for the whole dataset 37
4.5 Weekly aggregated cases of best models compared to ICOSARI data 38
4.6 Weekly aggregated cases of best models compared to no diagnosis codes available 39
4.7 Alluvial plot for the predicted diagnoses of DataWigBoW 40
A.1 Reference cases compared to ICOSARI data 56
A.2 Best models compared to ICOSARI data for the whole dataset 56
A.3 Recall by class frequency for DataWigBoW 57

List of Abbreviations

fn False negative

fp False positive

tn True negative

tp True positive

ARI Acute respiratory illness

BoW Bag of words

bpm Beats per minute

brpm Breaths per minute

ED Emergency department

ESEG *Erkennung und Sicherung epidemischer Gefahrenlagen*

GP General practitioner

ICD International Classification of Diseases

IfSG *Infektionsschutzgesetz*

ILI Influenza-like illness

LogReg Logistic regression

LSTM Long short-term memory

MAR Missing at random

MCAR Missing completely at random

mmHg Millimeters of mercury

MTS Manchester Triage System

NB Naive Bayes

NIM Non-ignorable missingness

NMAR Not missing at random

RKI Robert Koch Institute

SARI Severe acute respiratory illness

SD Standard deviation

SVM Support vector machine

SyS Syndromic surveillance

WHO World Health Organisation

1 Introduction and motivation

Emergency department data enable a near real-time surveillance of possible public health threats like natural disease outbreaks (Stoto et al., 2006; Henning, 2004). If they are available in electronic form, they can help to detect potential events earlier than would be possible with conventional surveillance systems. This is why emergency department data serve as an important data source for syndromic surveillance.

A real-time syndromic surveillance system is being piloted for Germany right now. As the main data source, emergency department (ED) data is routinely collected from several hospitals all over Germany. This dataset suffers from missing values in some variables, most importantly in the clinical diagnosis. The diagnosis, in turn, is a crucial basis for numerous analyses (e. g. outbreak detection) and needs to be available at a very early time point to enable the surveillance of diseases or other public health threats on hourly basis. Syndromes that are monitored in syndromic surveillance are usually based on expert case definitions that in turn are based on diagnosis codes.

Therefore, this work explores and evaluates two approaches of dealing with missing values in the diagnosis variable. It will thereby enhance the emergency department dataset and enable the surveillance of syndromes in a timely manner. An influenza-like illness syndrome will be used as an example throughout the analysis. The first approach will explore two methods to impute and predict missing diagnosis codes from the remaining mixed type variables in the dataset. In the second approach, two classification models are trained to predict the influenza-like illness syndrome without relying on the diagnosis variable.

In the introduction, the concepts of syndromic surveillance, influenza-like-illness and the existing surveillance systems in Germany are briefly presented. This is followed by an overview of related work on automatic coding systems that also try to predict diagnosis codes and on automatically finding case definitions for syndromes. Following this, the precise goals of the thesis are defined.

1.1 Syndromic surveillance

Syndromic surveillance (SyS) can be described as the systematic collection, analysis, and reporting of real-time health-related data in order to enable public health action (Triple S Project, 2011). Syndromic surveillance can make use of various data sources, including diagnostic as well as pre-diagnostic information, such as preliminary diagnoses or symptoms (Katz, May, Baker, & Test, 2011). It can serve different use cases, such as early warning and outbreak detection of diseases, monitoring of health as well as communicable and non-communicable diseases, or the evaluation of public health interventions. Syndromic surveillance systems have been successfully implemented in several countries with different goals. In the United Kingdom the system uses information from emergency departments, general practitioners (GPs) and a health service hotline to monitor seasonal respiratory infections (H. Hughes et al., 2016), evaluate a vaccination program (Bawa et al., 2015), and monitor non-communicable diseases in relation to extreme weather events like cold weather or heatwaves (Elliot et al., 2014; H. Hughes et al., 2014).[1] In Italy, a system using police reports together with health data was used to assess the impact of road traffic injuries (Chini et al., 2009). Zheng, Aitken, Muscatello, and Churches (2007) report the successful use of ED data to monitor influenza in Australia. In France, a syndromic surveillance was established in 2004, which uses data from emergency departments, general practitioners and mortality information to monitor the public

[1] For more details see ReSST (Real-time Syndromic Surveillance Team).

Glossary[3]

Diagnosis
the identification of a disease or injury made by a doctor
can be coded according to the International Classification of Diseases (ICD) manual
e.g. J22 ("Acute lower respiratory infection")

Symptom
an indication of a disorder or disease
e.g. pain, cough, fever

Syndrome
a group of symptoms that collectively indicate or characterize a disease
e.g. Influenza

Case definition
the rule or pattern defining whether or not a patient has a syndrome
e.g. ICD-diagnosis code is one of J09 - J22

health on a national and regional level (Caserio-Schönemann, Bousquet, Fouillet, & Henry, 2014).[2]

At the Robert Koch Institute (RKI), a real-time syndromic surveillance system is currently being piloted for Germany. Routinely collected emergency department data from several hospitals all over Germany serve as the main data source. They are available retrospectively from 2012 and the integration of new data on an hourly basis is currently implemented. The information included in the data is demographics of the patient, hospital administrative and health-related information (e.g. preliminary diagnoses, patient chief complaints, and vital parameters). With this new syndromic surveillance system being established in Germany, a further system can be used to detect potential public health threats. It can complement conventional established surveillance systems like lab-confirmed cases (as described in Section 1.2), and can furthermore enable an even earlier detection of public health threats than possible with the conventional systems.

To detect potential cases of certain diseases in SyS, one can either use diagnostic information like diagnosis codes or predefined groups of symptoms (i. e. syndromes) that are probably related to the manifestation of the diseases of interest. These variables differ in their availability in an ED setting. Whereas symptoms are most of the time available at a very early stage, diagnosis codes might be only available at the end of a stay or even missing completely. This can be due to several reasons. Symptoms have to be reported within the initial admission and triaging process, but diagnoses are mostly given at the end of a stay in the ED in order to conclude the case. They can be missing completely, if for example a patient is revisiting the ED for a change of dressing or consultation and is not getting assigned a new diagnosis. A diagnosis might also be missing if a case is not concluded, or because it has been forgotten to assign a clinical diagnosis.

Relying solely on the diagnosis code for detecting diseases or public health threats might therefore be impractical, especially when a large amount of diagnosis codes are missing or if the codes are not available at an early time point. Symptoms on the other hand can be available at a very early time point and can be used to assign visits to syndromes. For

[2]Information in French: SurSaUD (Surveillance Sanitaire des Urgences et des Décès).

[3]American Heritage® Dictionary of the English Language, Fifth Edition. (2011). Retrieved April 4 2020 from https://www.thefreedictionary.com/.

some surveillance objectives like infectious disease outbreaks, it has been shown that the surveillance of syndromes makes the detection of these events possible at an earlier stage (Stoto et al., 2006). Therefore, symptoms seem to be a suitable pre-diagnostic information to detect potential cases of a certain disease or syndrome. This is why in the present work symptoms are considered as the main informative source for predicting the health outcome syndrome, additional to diagnosis codes.

As an example for a syndrome that has to be monitored this work uses seasonal influenza (or more specific, an influenza-like illness syndrome). It was chosen because for this syndrome established case definitions and multiple surveillance systems already exist, which might serve as external sources of validation. The next section provides a brief overview of seasonal influenza, its surveillance and influenza-like illness.

1.2 Influenza and influenza-like illness

Seasonal influenza is an acute respiratory infection caused by influenza viruses (World Health Organization, 2018). Symptoms include the sudden onset of fever, cough, headache, muscle and joint pain, feeling unwell, sore throat and a runny nose. It is transmitted via infectious droplets and can cause severe illness or death for people at risk (World Health Organization, 2018). Seasonal epidemics occur mainly during the months January to March and can infect a large part of the population, causing high mortality with up to 72000 deaths a year in the European Region (WHO Regional Office for Europe, 2020).

A broader group of illnesses with the above mentioned symptoms are collectively called influenza-like illness (ILI). Infections caused by other respiratory viruses can cause the same symptoms and with influenza lacking specific symptoms, this can complicate the differentiation of influenza from other pathogens. But for some surveillance tasks the broader definition of ILI is used. The exact case definition of ILI varies across countries (Kalimeri et al., 2019; Casalegno et al., 2017; Aguilera et al., 2003; Jiang et al., 2015). To make a definitive diagnosis of influenza, a laboratory diagnostic test is required.

In Germany, the RKI, together with the Federal Ministry of Health and local health authorities, is responsible for the public health surveillance. The Infection Protection Act *(Infektionsschutzgesetz (IfSG))* regulates the notification of specific infectious diseases, which have to be reported to the health authorities by physicians and laboratories (Bundesamt für Justiz, 2000). These cases build the basis for a nationwide surveillance of infectious diseases, but only represent the laboratory confirmed cases. Several additional surveillance systems are used to complement this data basis and also capture less severe cases. They are more sensitive because they are not dependent on the confirmation by laboratories and therefore capture more cases. The Working Group Influenza *(Arbeitsgemeinschaft Influenza)* uses data from general practitioners for a nationwide surveillance of acute respiratory infections leading to a visit to the doctor (Robert Koch-Institut, 2020a). And the weekly online survey *GrippeWeb* provides a self-reporting tool for any voluntary participant to also include cases that are not necessarily accompanied by consulting a doctor (Robert Koch-Institut, 2020b). Additionally, the sentinel hospital surveillance *ICOSARI* uses information on hospitalised cases with influenza diagnoses from about 80 hospitals all over Germany to assess the seriousness of acute respiratory infections. This system also captures more severe cases, but does not require the diagnosis to be lab-confirmed (Buda, Tolksdorf, Schuler, Kuhlen, & Haas, 2017). At last, as a first electronic surveillance system of acute respiratory illnesses (ARI), the $SEED^{ARE}$ system uses electronically transmitted ARI cases by 70 sentinel practitioners (Köpke, Prahm, Buda, & Haas, 2016).

All of these systems lack in a certain timeliness. Some of them are paper-based and take 1-2 days up to two weeks for the data to be available at the RKI. The additional surveillance systems

are limited to only one disease and each of the systems monitors a slightly different group of diseases. The resulting case numbers are therefore not easily comparable. A syndromic surveillance system as proposed earlier would tackle these shortcomings in two ways. First, by providing information on real-time basis, a timely monitoring is possible. Second, with the data not being selected solely for one purpose and therefore pre-selected to a specific use case, any health indicator can be monitored. Researchers can define their own syndromes and different data sources can be combined.

As mentioned earlier, different systems use different types of syndrome case definitions due to different use cases. For instance, some case definitions are simple rule based decisions that use ICD codes or symptoms to define whether a case has the monitored syndrome or not (e.g. the discharge diagnosis has to be one of J09 - J22, as in the SARI-surveillance described by Buda et al. (2017)). Others try to automatically define the syndrome, either supervised (Olszewski, 2003; Espino et al., 2006) or unsupervised (Kalimeri et al., 2019), in which case they are often based on symptoms.

To define case definitions for the ED data, we have to consider that the final clinical diagnosis is often only available at a later time point or is missing completely, whereas symptoms are usually available right after admission (Krey, 2016). Thus, using rule based case definitions (e.g. for ILI) that heavily depend on the existence of ICD codes will not work well on the real-time data, because the cases with a missing diagnoses can not be categorised into the syndrome. To avoid losing significant information when creating case definitions, two approaches are presented in this thesis: first, imputing the missing diagnosis from the other available information and using the rule based definition to find ILI cases; and second, predicting an ILI syndrome based on the remaining information using a classification model that does not rely on the ICD code. Some work has been done in both of the fields which will be presented in the next two sections.

1.3 Automatic prediction of clinical diagnoses

Clinical diagnoses can be classified according to the International Classification of Diseases (ICD), published by the World Health Organization (2020). The current 10th version is available in a national German version and physicians and psychotherapists are obliged to code their diagnoses according to this system by law (Bundesamt für Justiz, 1988). The coding system has a mono-hierarchical structure that divides the diseases into chapters, groups, categories and subcategories. See Table 1.1 for an overview of all chapters. A code is alphanumerical and can have up to seven characters. The first character is a letter, indicating the chapter. The second two characters are numeric, indicating the category (or block). This is followed by a decimal point with the last digits giving more detailed information on the disease.

Example: J01.1

J:	Diseases of the respiratory system	(chapter/letter)
J00 - J06:	Acute upper respiratory infections	(group)
J01:	Acute sinusitis	(category/block)
J01.1:	Acute frontal sinusitis	(subcategory)

There is a large field of research on automatically predicting a clinical diagnosis code. The task is usually to use unstructured textual data (e.g. patient notes, discharge summaries, radiology reports) and some structured data (e.g. vital parameters, demographic information) to predict a diagnosis code (Stanfill, Williams, Fenton, Jenders, & Hersh, 2010; Scheurwegs, Cule, Luyckx, Luyten, & Daelemans, 2017). Because this can be seen as a (text) classification problem, the underlying methods are usually classification methods like support vector

Chapter	Letter	Block	Title
I	A, B	A00–B99	Certain infectious and parasitic diseases
II	C, D	C00–D48	Neoplasms
III	D	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E	E00–E90	Endocrine, nutritional and metabolic diseases
V	F	F00–F99	Mental and behavioural disorders
VI	G	G00–G99	Diseases of the nervous system
VII	H	H00–H59	Diseases of the eye and adnexa
VIII	H	H60–H95	Diseases of the ear and mastoid process
IX	I	I00–I99	Diseases of the circulatory system
X	J	J00–J99	Diseases of the respiratory system
XI	K	K00–K93	Diseases of the digestive system
XII	L	L00–L99	Diseases of the skin and subcutaneous tissue
XIII	M	M00–M99	Diseases of the musculoskeletal system and connective tissue
XIV	N	N00–N99	Diseases of the genitourinary system
XV	O	O00–O99	Pregnancy, childbirth and the puerperium
XVI	P	P00–P96	Certain conditions originating in the perinatal period
XVII	Q	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S, T	S00–T98	Injury, poisoning and certain other consequences of external causes
XX	V, W, X, Y	V01–Y98	External causes of morbidity and mortality
XXI	Z	Z00–Z99	Factors influencing health status and contact with health services
XXII	U	U00–U99	Codes for special purposes

Table 1.1: ICD-10 chapters. For each chapter, the number, the corresponding letter, the block (code range) and the title from the international version of the ICD-10 is shown. Adapted from World Health Organization (2020).

machines (SVM) (Baumel, Nassour-Kassis, Cohen, Elhadad, & Elhadad, 2018; Perotte et al., 2014), naive Bayes (Pakhomov, Buntrock, & Chute, 2006; Scheurwegs et al., 2017), or random forest (Scheurwegs et al., 2017). The resulting systems can help medical staff to assign codes by making suggestions in a computer-assisted coding way (Pakhomov et al., 2006; Larkey & Croft, 1995), therefore automating a usually expensive and inefficient coding process (Stanfill et al., 2010).

Some approaches try to incorporate the hierarchical structure of the classification system into their models or evaluation metrics. Perotte et al. (2014), for example, use the tree-like hierarchy of ICD-9 codes in the training of several SVM classifiers, with one classifier for each code.[4] In the training, only those cases are included where the parent node for this code is positive. In testing, the classifiers are applied downwards from the root until one child is predicted negative, resulting in a subtree of multi-label predictions for a given case (Perotte et al., 2014). They also define alternative evaluation metrics that consider the hierarchical nature of predictions and may give a better evaluation of this task. Taking into account the hierarchical structure improved the performance from 27.6% to 39.5% in F-measure (Perotte et al., 2014).

It can be seen that some approaches exist for predicting diagnosis codes from structured and unstructured data to help making decisions for single cases. Taking into account the hierarchical structure of the used diagnosis system can improve the performance (Perotte et al., 2014). Most of the systems described here are designed to suggest potential diagnosis codes for a specific single case, with a medical trained person having to make the final decision. In this work, the initial data differ slightly from those used in the described studies, because for example patient records are not available. Therefore, the focus is on imputing diagnoses to have a better overall impression and not to make the best prediction for a single case. Once the diagnoses codes are predicted, further analyses are possible.

1.4 Syndrome prediction and unsupervised learning of case definitions

Some work has been done on predicting a syndrome and automatically finding a case definition without relying on the diagnosis code. Kalimeri et al. (2019) propose an unsupervised framework where they use non-negative matrix factorisation to find clusters of symptoms in a time series of self-reported symptoms, which can explain the presence of ILI. The advantage of this data-driven approach is that relevant symptoms do not need to be defined beforehand but result from the data. This enhances the flexibility of case definitions and allows the consideration of country-specific characteristics. At the same time this might complicate the comparison between different data sources. Moreover, self-reported symptoms can be biased by the reporting behaviour of the participants or suffer from self-selection.

Other works use patient chief complaints that are reported in emergency departments to develop syndromes for syndromic surveillance. Sniegoski (2004) proposes an approach, where chief complaints in textual form are automatically classified into syndrome groups and therefore replace a time consuming manual classification. The work of Espino et al. (2006) resembles this approach, they also use chief complaints in textual form to first predict specific symptoms and in a next step classify them into syndromes. Olszewski (2003) compared the predictive use of textual symptoms against categorical ICD-based symptom codes for classifying patients into syndromes. Here the free-text symptoms were better predictors than ICD-9 symptom codes.

[4] ICD-9 codes have a hierarchical structure as well, consisting of 3 to 5 digits. See DIMDI: ICD-9 - Internationale Klassifikation der Krankheiten for more details.

Apart from the unsupervised approach by Kalimeri et al. (2019) all of the studies used predefined syndromes with manually labelled training data based on more or less clear rules. While this might lead to more comprehensible models, it still involves a great amount of human input with hundred thousands of visits that need to be classified by hand before. Therefore, it could be less expensive to use an unsupervised approach, or to use already existing syndrome definitions based on clearly defined rules as a label, to then train a model which subsequently is able to automatically classify the syndrome for new unseen data.

1.5 Goals of this thesis

This thesis is aimed at enhancing emergency department data to enable a near real-time syndromic surveillance. As a use case for this work, the prediction of an influenza-like illness syndrome is investigated. The so far standard to find cases that might have an ILI syndrome is to use a rule-based definition that is developed by epidemiologists. This definition mainly relies on the existence of ICD-10 diagnosis codes. However, this data suffers from missing values in the clinical diagnosis variable. In about 16% of the visits this information is missing. The reasons for this can vary and are explained in Section 1.1. When the diagnosis code is missing, the expert-defined rule is not applicable and the estimated amount of relevant cases with the syndrome is likely to be incomplete. Additionally, for a time range of about four years (2012 - 2016), no diagnosis codes are available at all due to a system error. This makes analyses based on ICD codes within the above time interval impossible. Because in the future more emergency departments will be included in this new syndromic surveillance system, it can be expected that they will suffer from missing values in this variable as well, possibly also to a higher extend.

Therefore, the present study investigates two approaches for dealing with the missing values in the diagnosis variable. As a use case, the goal is to classify the existence of an influenza-like illness syndrome for a given ED visit. Cases that might have the ILI syndrome ought to be detected without relying on the diagnosis code but solely on other information available.

As shown in Section 1.3, there is substantial evidence in the literature that clinical diagnoses can be predicted from other health related data, as is available in a dataset of emergency department data. Therefore, the first approach explores two methods to impute and predict missing diagnosis codes based on the information from the other variables in the dataset. These imputed diagnosis codes can then be used to apply the rule-based definition of the ILI syndrome. Subsequently, the quantity and seasonality of this syndrome in the given time frame can be assessed. The two methods that are compared in their performance of predicting ICD-10 diagnosis codes are: a multi-class naive Bayes classifier and a machine learning approach, that extracts information from categorical, numeric and textual variables to impute missing values. Additionally, a naive Bayes model is implemented that takes into account the hierarchical structure of ICD-10 codes.

In the second approach, the health outcome influenza-like illness is predicted directly from the data, therefore obviating the need for a predefined set of rules. The works presented in Section 1.4 suggest that syndromes, which are generally used in syndromic surveillance or triaging situations, can be automatically generated from symptom variables. The goal is to extend human-made rule sets by learning from data (as in the unsupervised approach by Kalimeri et al. (2019)) and replace the time consuming manual classification. To label the training data, the already existing expert case definition of the ILI syndrome is used. The classification model will then be based solely on the symptom variables and all other information available in the dataset, except for the diagnosis code. This way, a case definition is automatically created that is independent of the existence of the clinical diagnosis. For this

binary classification task, a logistic regression model is compared to a naive Bayes classifier.

Because case definitions are usually rule based (combinations of diagnoses and symptoms), it will be important to provide explanations for the new definitions that are based on statistical models in order for epidemiologists to accept this approach. This work therefore aims at making the decision mechanisms of the models comprehensible.

The quality of the imputation of diagnoses as well as of the prediction of the syndrome is evaluated for each model. It is mainly of interest whether the models detect as many ILI cases as the rule based definition and how the predictions differ from those cases resulting from the expert rule. A good model should find as many ILI cases as possible from the originally present ILI cases. As a last step, the models are evaluated on a time series level. For this, the resulting ILI cases are aggregated on a weekly basis and compared to each other in terms of seasonal peaks and total amount of cases. Additionally, they are compared to an external data source that approximately shows the same scenario of seasonal ILI cases.

The thesis is divided into the following parts. First, the dataset used in this work is described in detail. This includes the description of the variables and their missingness, a description of the preprocessing steps, an exploratory data analysis and the implementation of the expert case definition of the ILI syndrome. Additionally, several external data sources are compared to find a suitable validation set. This is followed by a methodological introduction to the concepts of imputation and classification and a description of the models and evaluation metrics used in this work. Following this, the analysis procedure is derived and its implementation is described in detail. The results for all experiments are presented and evaluated, starting with the imputation of the diagnosis codes, followed by the prediction of the ILI syndrome and eventually the evaluation on time series basis. Last, the results of the analysis are discussed. The limitations of this work are considered and a conclusion is drawn.

2 Description of the dataset

In this section, the dataset used in this work is presented in detail. After explaining the origin of the data, the variables in the dataset are presented and the preprocessing steps explained. Within the exploratory data analysis, the problem of missing values in the diagnosis variable and in other variables becomes visible. The frequencies and distributions of the variables used in the analyses are shown and described. At last, the expert definition of the ILI syndrome is explained, which is used as the internal reference throughout the analyses, and an additional data source is chosen as the external reference in the analysis.

2.1 Data source

Within the ESEG project *(Erkennung und Sicherung epidemischer Gefahrenlagen)*[5] a group of emergency departments provide retrospective data. In the future, near real-time data will be made available additionally. For this analysis, data from only one ED is used. It contains routinely collected information on all patient visits between February 2012 and June 2019, a total of n = 384021 visits. Because the data is anonymised, one patient visiting the ED multiple times can not be identified as the same patient. Therefore each observation represents a visit and not a patient.

2.2 Variables and preprocessing

The dataset consists of 16 different variables which can be grouped into three different domains. The first domain contains demographic information on the patient like age and gender. The second domain contains administrative details like time and date, referral (whether a patient was send to the ED by a general practitioner, came by himself or with an ambulance etc.) and the department involved in the case. The third domain contains information on the disease like the chief complaints and diagnosis values as well as some vital parameters (heart rate, temperature, systolic blood pressure, respiratory rate and oxygen saturation). Some other variables are also available but were excluded from the analysis because of a high amount of missing values or because they were irrelevant for the task (e.g. vaccination status, postal code). All variables that were originally included in the dataset together with their amount of missing values can be found in Table A.1 in Appendix A. Note that the variable names are given in English, but their values are recorded in German only.

Complaint (group)	Complaint (value)
Collapsed adult	recent problem ($<$ 7 days)
Falls	recent light pain (VAS 1-3, $<$ 7 days)
Falls	uncontrolled small bleeding
Back pain	moderate pain (VAS 4-7)

Table 2.1: Examples for the variables complaint group and complaint value.

Patient chief complaints are encoded using the Manchester Triage System (Krey, 2016), a coding system that has a hierarchical structure. The first applicable and most severe condition is taken as the major complaint, consisting of both a complaint group (51 possible categories) and a value (191 possible indicators). Therefore, the two categorical variables

[5] See RKI: Detection and Protection of epidemic situations (ESEG) for more details.

complaint group and complaint value form the chief complaint of a visit. Several group-value pairs can be present in one visit due to updates after re-triaging of the patient, so the last recorded value is seen as the most valid one and therefore chosen for the analysis. Examples for complaint group and value pairs are shown in Table 2.1.

Additionally to the MTS patient chief complaints, **clinical diagnosis codes** are assigned for a visit. They are encoded by the ICD-10-GM (German version) (World Health Organization, 2020). Four different diagnosis categories are possible: G (*Gesichert*, verified), V (*Verdacht*, suspected), A (*Ausschluss*, excluded) and Z (*Zustand nach*, condition after). Only the verified "G" diagnosis is used in the analyses. One visit can get assigned several diagnosis codes. Unlike in the MTS complaints, where only the last value is likely to be the valid one, several diagnosis codes can be valid. Because we do not know which diagnoses are most relevant for one visit, only the first recorded diagnosis code is used for the analysis. It has to be taken into consideration that this might not be the most relevant diagnosis or that others might be equally relevant. Nevertheless this first diagnosis is considered as the reference value for the rest of the analysis and used as the target for the classification tasks.

As described in Section 1.3, an ICD-10 diagnosis code consists of up to five digits. Due to anonymisation of the dataset, the second decimal point (last digit) is already deleted before the data transfer. In this analysis, two simplified versions of the ICD diagnosis code were extracted to reduce the amount of possible classes. This procedure follows the work of Subotin and Davis (2014) and Baumel et al. (2018) on classification of ICD codes, who also use truncated versions of the diagnosis codes. Because of the hierarchical nature of ICD-10 codes (as described in Section 1.3), the higher level can be extracted easily. In this case, from a given diagnosis (e.g. A09.0), the category (or block) was saved as diagnosis (block) (e.g. A09) and the overall chapter (first letter only, e.g. A) was saved as diagnosis (letter). This resulted in 1135 possible diagnosis codes instead of 3741 for diagnosis (block).

The variable **department** can also take on several values for one visit, so the first value was selected for analysis as well. **Gender**, **age** and **referral** could be used without further preprocessing. For all **vital parameters**, which can take on several values for one visit, the median of all occurring values for one visit was computed and used in the analysis. Other forms of feature extraction are possible as well and can be considered in follow-up work. From the **date** and **time** information, corresponding **weekdays** and **months** are extracted and are used as a feature in the analysis as well.

This preprocessing resulted in a dataset with one row per visit and maximum one value per variable per visit in accordance with the tidy data framework by Wickham (2014). The reduction of the dataset due to missing values in some of the variables is described in the next section.

2.3 Exploratory data analysis

During the initial exploration of the dataset and especially when examining the amount of missing values, the following characteristic became visible: The variable diagnosis, which is relevant for all of the planned analyses, is missing completely in the first half of the dataset, up until 2016-03-16. Figure 2.1 gives an impression of this pattern, which only occurs for this variable. After this date, the variable is missing with a moderate but still substantial amount.

Therefore the dataset was split in two parts, the first half including cases that occurred before 2016-03-16, the second part starting from this date up until June 2019. All analyses were conducted with the second half of the dataset. For a detailed description of dataset sizes see Table 3.1 in Section 3.3.1. The following descriptive statistics refer to the second half.

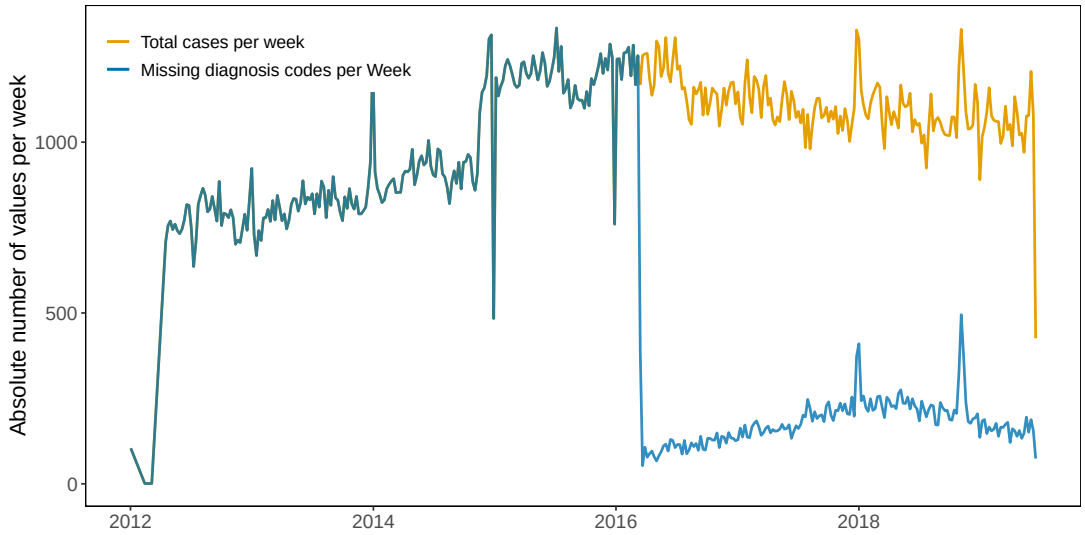


Figure 2.1: Absolute number of missing diagnosis codes vs. total cases per week. Both are shown for the whole dataset, from 2012 to 2019. Notice that in the first half both are exactly the same, showing that diagnosis codes are missing in all cases.

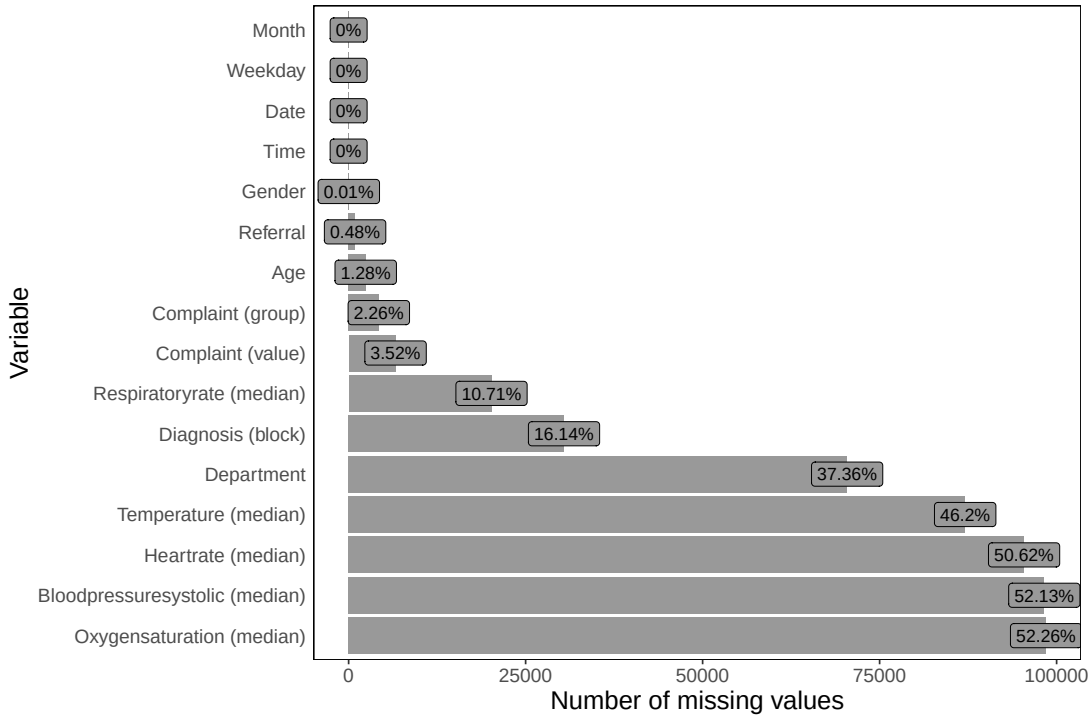


Figure 2.2: Percentage of missing values in each variable. The grey bars show the absolute number and the labels at the end of each bar the percentages of missing values for each variable in the second half of the dataset.

Missingness Figure 2.2 shows the total and relative amount of missing values in every variable. There are no missings in the date and hour variables, due to the fact that they are set by the ED system automatically. Gender, age, referral and complaints are available for almost all visits, missing only in less than 5% of the visits. Table A.3 in Appendix A shows the abbreviations used in the variable referral. The diagnosis variable is missing in 16.14% of the visits, the reasons for this were discussed in detail in Section 1. The department is missing in 37.36% of the visits. Vital values are missing in about half of the visits, except for respiratory

rate, which is missing in 10.7%.

To get a dataset that contains no missing values in order to train models (that can not deal with missing values), the second half of the dataset was reduced again and all rows containing at least one missing value in a variable except the diagnosis variable were deleted (complete-case analysis). The resulting dataset has a size of n = 47088. The following description of the dataset is based on this reduced dataset with no missing values.

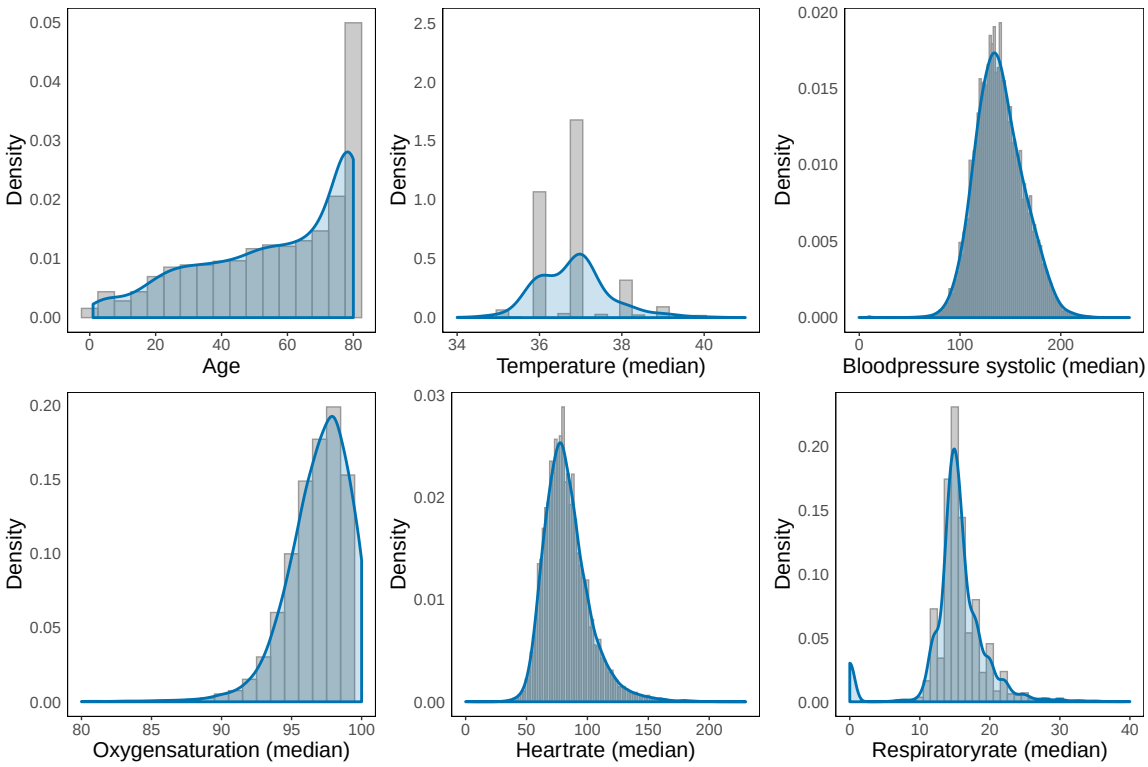


Figure 2.3: Binned frequencies and probability density of all metric variables. Shown for the variables age, temperature (median), systolic bloodpressure (median), oxygensaturation (median), heart rate (median) and respiratory rate (median) (from top left to bottom right).

Descriptive statistics 50.97% of the visits are female and the age mean is 56.32 years (SD = 22.49). Having a closer look at the first plot in Figure 2.3 we can see that it is more the elderly that are admitted at the ED.

Figure 2.4 shows frequencies for all categorical variables. The letter (chapter) "R" of diagnosis (letter) which is "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" is the most frequent class in the ED. This chapter contains symptoms rather than diagnoses, that physicians can use if no specific diagnosis can be assigned, but a code is needed for administrative purposes. The second most frequent diagnosis (letter) is "I" ("Diseases of the circulatory system"). The most frequent involved departments are internal medicine, neurology, surgery and pediatrics. Most patients are not referred to the ED by a physician (*NPHYS*), meaning that they come by themselves. Some arrive with the ambulance (*RD*) and some are send by a general practitioner (*VAP*). The visits follow a regular daily pattern with the most patients arriving in daytime, with a peak in the morning and less people coming in at night. Regarding the weekday only a slight increase in visits can be observed for Mondays and Thursdays and the fewest visits on weekends. Over the time of the year, peaks are visible in May, June and March and less cases in autumn.

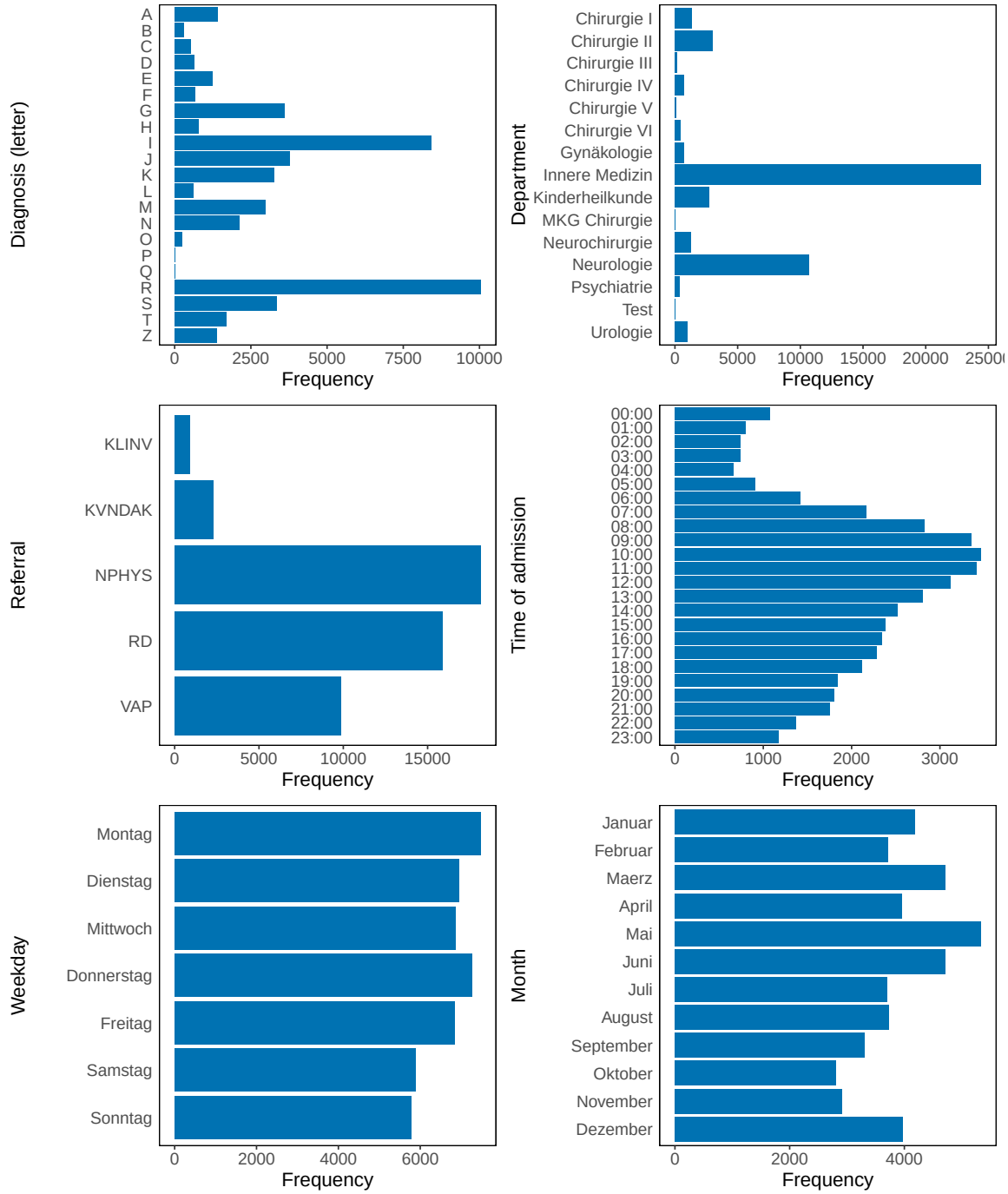


Figure 2.4: Frequencies for all categorical variables. Absolute number of occurrences are displayed for the variables diagnosis (letter), department, referral, time of admission, weekday and month (from top left to bottom right).

Figure 2.3 shows binned frequencies of the metric variables together with the densities. Body temperature has a mean of 36.80°C and SD = 0.92°C. Systolic blood pressure follows a normal distribution with a mean of 139.50 mmHg and SD = 23.95 mmHg. Oxygen saturation as well follows a normal distribution with a cut at 100% (mean = 97.07%, SD = 3.15%). Heart rate has a positive skew with mean = 85.61 bpm and SD = 23.00 bpm. Respiratory rate has a mean of 13.23 brpm and SD = 6.30 brpm, with a few cases having a respiratory rate of 0. This can be due to a stop of breathing, the death of a person or a reporting or system error.

A comparison of the mean and standard deviation of the metric variables for the different datasets is shown in Table A.2 in Appendix A. It can be seen that the vital parameters do not change much in the different datasets, but the dataset containing only complete cases (dataset *(e)*) has less younger people and therefore a higher mean age (56.32 years versus 45.51 in the whole set). This might be an indicator that variables like department, referral or vital parameters are missing more in children. The proportion of male and female patients is almost the same for all datasets, with the last dataset having slightly more females. A comparison of the categorical variables showed that the most frequent diagnosis codes (letter) are S ("Injury, poisoning and certain other consequences of external causes") and R for the whole dataset compared to R and I in the reduced dataset. Additionally, there are two peaks of visits in the course of a day in the whole dataset, one in the morning and one in the afternoon. In the reduced dataset, there is only one peak in the morning. These findings might indicate that values are missing more in certain conditions and have to be considered when generalising the results of the models.

2.4 Expert case definition of the ILI syndrome

As an initial definition of the ILI syndrome, the expert case definition made by an epidemiologist at the RKI is used. Even though other definitions might be possible and every definition will result in slightly different groups of cases, this one is used as the reference inside this work. It uses simple rules based on ICD-10 diagnosis codes, MTS complaints as well as the body temperature measured for the patient to define whether the patient might have ILI. An example rule, based solely on diagnosis codes, is

```
diagnosis = 'J09' OR diagnosis = 'J10' OR diagnosis = 'J11'.
```

Another rule, that is based on MTS complaints and body temperature is given by

```
temperature >= 38 AND
((complaint_group = 'Atemnot bei Erwachsenen AND
  complaint_value = 'Infektion der Atemwege') OR
 (complaint_group = 'Atemnot bei Kindern' AND
  complaint_value = 'Infektion der Atemwege')).
```

The case definitions vary in their specificity and sensitivity. They can be chosen depending on the use case and will result in different amount of cases. In this work, a sensitive case definition was chosen, in order to obtain many positive examples for training. To achieve this, the different expert rules were combined in a way that if at least one rule applies to a case, it is labelled with ILI. This labelling results in an additional column in the dataset, syndrome ILI, indicating if a visit might have the ILI syndrome according to the expert rules or not (type: logical).

Some of the rules are based on the original 4-digit diagnosis (e.g. J09.0). Because only the 3-digit diagnosis (block) is used in this analysis, it had to be ruled out that this would change

the meaning of the definitions. A consultation with the epidemiologists revealed that only a few 4-digit diagnosis categories exist, that would falsely be labelled as ILI. A descriptive analysis showed that this does concern 12 cases in this dataset ($N_{ILI} = 1401$ vs. $N_{ILI\_block} = 1413$), which was considered a negligible amount of cases. Therefore the ILI-expert rule was applied to the 3-digit version (diagnosis (block)), but it has to be kept in mind that this procedure might not be transferable to another dataset or case definition.

2.5 Comparison with external data sources

Several other surveillance systems for influenza, influenza-like illness or acute respiratory infection were introduced in Section 1.2. Some of them were considered to serve as an external reference for validating the models, but each of them uses a different case definition for the monitored health outcome. Figure 2.5 shows the amount of cases of each external data source (relative to its maximum value) compared to the influenza-like illness cases obtained by the expert rule described in the last section (dark blue line), which serves as the internal reference for all models in this work.

The cases from the traditional surveillance *(IfSG-Meldedaten)* contain only laboratory confirmed influenza cases. They are shown for the same federal state that the hospital of this study lies in (yellow line) and for the same district (grey line). The lab-confirmed cases are a very specific case definition of influenza and therefore result in very distinctive peaks within the influenza season, with almost no cases outside. Nevertheless, influenza cases can occur also in other times of the year, but the amount of testing is higher during the seasons (and varies over the years). The IfSG-data have correlations of r = 0.660 ($p < 0.001$, federal state) and 0.691 ($p < 0.001$, district) with the internal reference.

The sentinel surveillance data of acute respiratory infections ($SEED^{ARE}$) contains cases reported by sentinel general practitioners, that got assigned a diagnosis code of either J00 - J22, J44.0 or B34.9 (Köpke et al., 2016). This includes not only influenza cases, but all acute respiratory infection cases, therefore resulting in a wider time series of weekly cases (see the light blue line in Figure 2.5, called ARE). It has a correlation of r = 0.729 ($p < 0.001$) with the internal reference, but diverges in the off-season. This might also be due to the restriction of the dataset in this study, which uses only data from one emergency department.

More severe acute respiratory illness cases are monitored via the ICOSARI system, using cases provided by sentinel hospitals about hospitalised cases with the diagnosis J09 - J22 (Buda et al., 2017). The resulting time series of weekly cases resembles the ARE cases, but has less deviations in between the seasonal peaks (see the red line in Figure 2.5). These cases have a correlation of r = 0.786 ($p < 0.001$) with the internal reference, which is the highest correlation of an external data source with the ILI cases from this study. The sentinel SARI cases were therefore chosen as the external reference, and it can also be assumed that the case definition resembles the one used in this study most.

In the first half of the dataset, no diagnosis codes are available. Therefore, the ILI case definition from this study can not be applied to obtain ILI cases. To still be able to assess the performance of the models developed in this study, they can be compared to the ICOSARI data in the first half. It is assumed that this external data source might provide an appropriate source of reference in this case, because it resembles the internal reference well in the second half of the dataset. For a plot with both the internal and external reference cases for the whole time frame, see Figure A.1 in the Appendix A.

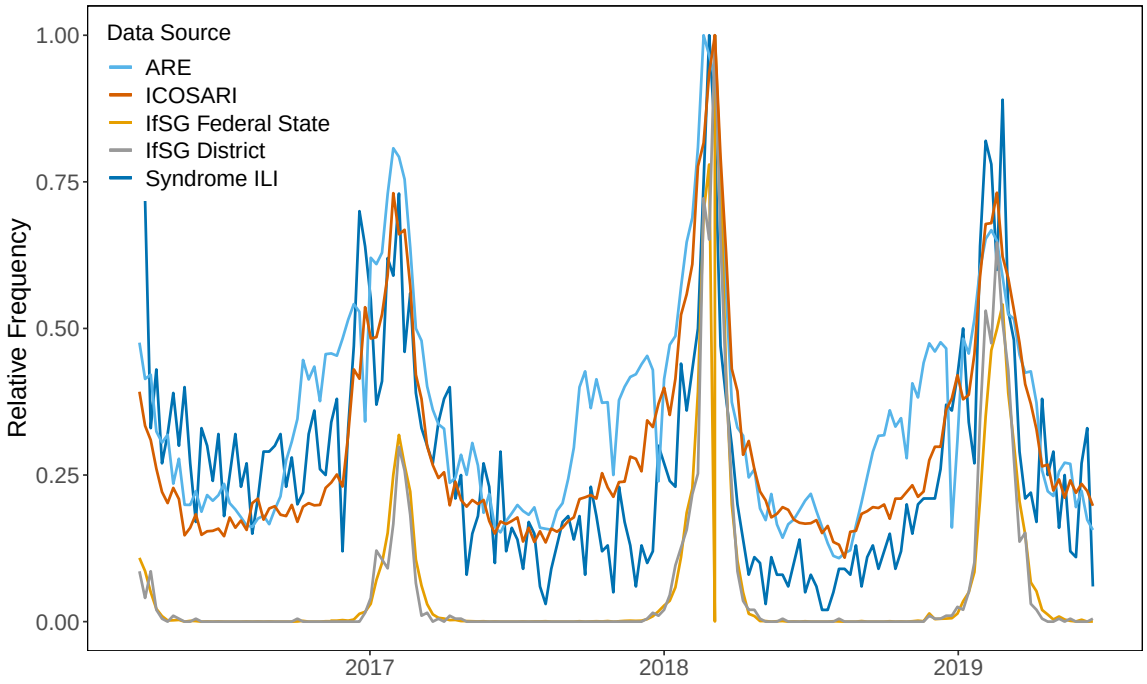


Figure 2.5: Comparison of time series from different external data sources. For several influenza and acute respiratory illness surveillance systems the relative amount of weekly aggregated cases are shown. To make the different data sources comparable, the cases are scaled to the maximum of each timeseries. Additionally, the expert defined ILI cases (internal reference) are shown (dark blue line).

3 Methods and implementation

This section presents the methods used in this work and their implementation in order to answer the research questions. In the first part of the work, the focus lies on imputing missing values. An introduction to the problem of missing values and the most commonly used methods is given. The methods used in this work are derived and explained.

The second part of this work focuses on predicting the ILI syndrome. Therefore a short introduction to the classification setting is given, followed by an explanation of the naive Bayes classifier and logistic regression. The problem of imbalanced datasets is addressed and up- and downsampling solutions are presented. It is explained how a classification method can be evaluated and the metrics used in this work are shown.

At last, the analysis procedure is explained in detail. The implementation of the different approaches is explained and the packages used for data analysis are named.

3.1 Imputation of missing data

When missing values occur in one or more variables, several approaches exist for dealing with them. Following the work of Little and Rubin (2019), the most common methods and their characteristics are shortly presented. The two approaches explored in this work are derived and explained.

3.1.1 Introduction to common imputation methods

The easiest solution for dealing with missing values in a variable is to assign a special symbol (e.g. "not answered") for a missing value in order to get a pipeline running. Most statistical methods will still need the user to exclude these cases from the analysis, but the information of how many cases are missing is preserved. If there is just a small amount of missing data, one can omit all rows with missing values (i.e. complete-case analysis, listwise deletion) and apply their methods. While this might be an easy and convenient solution for situations with very small amounts of missing data, it might lead to biases if a large part of the data has to be excluded or if the data is missing not at random. Then drawing inferences about the target population might not be appropriate anymore. The third option would be to impute the missing values, i.e. replacing the missing value in one variable for a given case with a probable other value for this variable and case.

Imputation methods that are commonly used are presented shortly. In hot deck and cold deck imputation, missing values are replaced with a constant value that was observed in similar responding cases in the same sample (hot deck) or an external source (cold deck). In mean imputation, missing values are replaced with the mean of observed values in the same variable (and the mode for non-numerical variables). Lastly, in regression imputation missing values in one variable are replaced with a predicted value from a regression of the missing variable on other variables that were observed for this case (e.g. MICE, Buuren and Groothuis-Oudshoorn (2010)). In all imputation methods, single and multiple imputations are possible. Multiple imputations result in more than one prediction for the imputed value and is therefore taking into account the uncertainty of this prediction, but it also results in higher computational efforts (Donders, Van Der Heijden, Stijnen, & Moons, 2006; Little & Rubin, 2019).

The mechanism that leads to missing data can have an impact on the performance of the different methods. In the best case, the missingness is independent of the data values, so the cases are missing completely at random (MCAR). If the missingness only depends on the other observed values and not on the values in the variable itself, the cases are missing at random (MAR). In the latter case, regression imputation was shown to still lead to unbiased

estimations (Donders et al., 2006). The last case is the one where the distribution of the missing values depends on the missing values themselves (not missing at random, NMAR) (Little & Rubin, 2019).

Most of the work on imputation methods has been done for imputing numerical values (Sentas & Angelis, 2006), but the variable that has to be imputed in this work (diagnosis) is categorical. When values are missing in categorical variables, a simple solution is to use the mode to impute the values, but this might lead to unsatisfying results. Another solution is to use a method in the consecutive analysis that is able to deal with missing data, like random forest or naive Bayes classification. Kalousis and Hilario (2000) for example compared several machine learning methods with regard to their tolerance of incomplete data and a naive Bayes classifier showed to be the most tolerant in their study. But this imputation problem can also be seen as a classification task and a classification model can be used to predict categorical missing data. Sentas and Angelis (2006) compared several imputation methods for categorical software data and found that multinomial logistic regression as a classification approach outperformed the other methods (listwise deletion, mean imputation, expectation maximization and regression imputation), even with a high number of values and at different missing mechanisms (MCAR, MAR and NIM (non-ignorable missingness)).

3.1.2 Classification approach

With this particular dataset, the to be imputed variable (diagnosis code) is categorical and the remaining input variables are numeric and categorical. Because the target variable (diagnosis) has to be existent for each case for further analyses, a complete-case analysis is not the right approach here. And with mode or hot deck imputation the highly differentiated variable would loose their informative power. Because other studies have shown that systems exist for predicting the diagnosis variable from other health related data (see Section 1.3), a classification approach could be used for imputing the missing diagnosis codes. The naive Bayes classifier (which will be presented in detail in Section 3.2.2) has proven to be well suited for categorical and numerical input data (Pakhomov et al., 2006; Scheurwegs et al., 2017) and has been used for the imputation of categorical data as well (Zhang, Kambhampati, Davis, Goode, & Cleland, 2012; Garcia & Hruschka, 2005). For this reason, it was used as a classification method for predicting the diagnosis codes in this work.

3.1.3 DataWig approach

As a second approach for imputing the missing diagnosis codes the package DataWig by Biessmann, Salinas, Schelter, Schmidt, and Lange (2018) is explored. They propose an end-to-end framework to impute missing values from numeric, categorical and sequential input variables. Biessmann et al. (2018) developed this package to enable the simple usage of their complex imputation models for different data types and with different neural network classifiers. For a given dataset with both numeric, categorical and sequential data, the user defines which variable is to be imputed and which are the possibly relevant input variables. The data type of the input variables can be determined by the algorithm using heuristics or explicitly stated by the user. The model itself proceeds with first encoding the data into numerical representations, then extracting relevant features and third learning classification models. At first the non-numerical data is encoded into a numerical representation. Categorical data is one-hot encoded (= transformation into dummy variables) and sequential data is transformed into a numerical vector (for details see Biessmann et al. (2018)). Depending on the type of each column, different featurizers are used in the next step to extract features from the numerical representations. For categorical columns, a one-hot encoded embedding layer is used

(see Biessmann et al. (2018)), for sequential columns either n-gram representations (i.e. bag of words approach) or long short-term memory neural networks can be used (LSTM approach, see Hochreiter and Schmidhuber (1997), Biessmann et al. (2018)). The extracted features are used together with the observed y values in the target column to learn to predict the target values in a supervised learning manner. Hyperparameter optimisation (finding a set of optimal hyperparameters for the algorithm) is done automatically by the model. Once the model is trained, it takes the input columns to make a prediction of the to be imputed class for every observation. The result is a likelihood for each possible y value in the target column given the input columns and the imputation model for every observation. The most likely value is taken as the imputed value.

The DataWig package was chosen as one of the imputation approaches because it is an already developed method that deals flexibly with mixed data types and enables the use of sequential variables as input columns. Because it can be assumed that the variable complaint contains the most information for the prediction of the diagnosis code, the focus when using the DataWig package will be on applying the different encoding and featurizing methods to this variable and compare the results. Here the variable complaint is in fact a categorical variable, because a predefined and limited number of answers are possible. But they can also be seen as textual descriptions of the complaint because they often appear as several words. They may therefore contain even more information for the prediction of the diagnosis code than when taking them as nominal. If similar complaints can be represented near to each other in the feature space because of a similar wording, this feature representation might contain more information than the categorical version of the complaints. Additionally, less frequent complaints might be better representable when they are close to other similar complaints, therefore enabling a better use of these complaints as well. Therefore, the encoding and featurizing methods designed for sequential variables are used as well. They are applied to the complaint variable to explore whether additional predictive accuracy will be gained. Furthermore, it might be possible that additional data sources will be available in future, possibly also containing free-text variables. Then this package would enable the inclusion of these additional variables in the model without developing a new imputation approach including text-classification methods.

3.2 Binary and multi-class classification

In this section, a brief introduction to the classification setting with binary and multi-class outcomes is given. The two classifiers, naive Bayes and logistic regression, that are used in this work, are shortly presented. The problem of unbalanced classes is addressed and the measures used for evaluating the models are described.

3.2.1 The classification setting

In a classification setting a categorical output variable Y and one or more input variables (or predictors) $X_1, ..., X_p$ are observed. For an assumed relationship $Y = f(X) + \epsilon$, with ϵ being an error term, f is estimated on the basis of training observations $(x_1, y_1), ..., (x_n, y_n)$, where $y_1, ..., y_n$ are qualitative (James, Witten, Hastie, & Tibshirani, 2013). The quality of this estimate can be quantified by the training error rate, as measured by the accuracy value ($\frac{tp+tn}{tp+tn+fp+fn}$), which equals to the proportion of correctly classified cases to all cases when the model is used to predict the training observations.[6] To compare several models and to assess how well the model performs on unseen data, the model is applied to a test set of unseen data

[6]For an explanation of tp, tn, fp and fn see Figure 3.1.

and the predicted class labels are compared to the actual labels, thus computing an error rate. This test error should be small for a classification model to be good (James et al., 2013).

In classification, the target variable can be binary, having two possible classes, or it can be a multi-class variable. In the multi-class setting, one can distinguish between a multi-label and a multinomial approach. In the former, each observation is assigned one or more labels, e.g. a probability for each possible label. In the latter, the classes are mutually exclusive and each observation gets assigned exactly one class.

Two classifiers that can be used when the predictors contain both numeric and categorical variables are logistic regression and the naive Bayes classifier. Both of them also provide an easily interpretable model and thus give insights to how the predictions are derived. They are shortly presented in the next two sections.

3.2.2 Naive Bayes

Given an input vector x, the naive Bayes classifier returns from all possible output classes $y \in Y$ the class $\hat{y}$ that has the maximum a posteriori probability given the input.

$$\hat{y} = \arg\max_{y \in Y} P(y|x) \tag{3.1}$$

Applying the Bayes rule and the simplifying assumption, that the probabilities $P(x_i|y)$ are independent given the class y, the following term results, describing the naive Bayes classifier:

$$c_{NB} = \hat{y} = \arg\max_{y \in Y} P(y) \prod_{x \in X} P(x|y) \tag{3.2}$$

The prior probability $P(y)$ of a given class y can be estimated easily by the frequency with which each output value y_c occurs in the training data. The likelihood is estimated using a multinomial approach and for metric input variables an underlying normal distribution is assumed. To prevent probabilities of exactly 0 (if a class and a feature never occur together in the training data), a Laplace smoothing can be used, which adds a small-sample correction of 1 so that no probability is ever exactly zero.

Even though the simplifying assumption of conditional independence is often not met, the naive Bayes classifier proves to still work well and even compete with more sophisticated models (Rish, 2001). It can be used both for binary as well as multi-class target variables and will therefore be used as a baseline approach for both tasks (imputation of diagnosis codes and prediction of an ILI syndrome) in this work.

3.2.3 Logistic regression

When the outcome variable of a prediction setting is binary, a logistic regression model can be used to model the relation. It models the probability that Y belongs to a certain category (James et al., 2013). The model formula can be written as

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{3.3}$$

where $X = (X_1, ..., X_p)$ are the p predictors. (3.3) can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{3.4}$$

This leads to the interpretation, that an increase of one unit in X_1 will change the log odds (the left-hand side in (3.3)) by β_1. The estimates for a logistic regression are obtained by maximum likelihood. A logistic regression model can take both numeric as well as categorical predictors, latter will be treated as dummy variables.

3.2.4 Imbalancedness

In many classification settings, the class one wants to detect is naturally under-represented (e.g. predicting spam vs. non-spam e-mails). A classifier that is always predicting the majority class will have a high accuracy, but will be poor at detecting the under-represented class. One solution is the up- or downsampling of the training dataset with regard to the class of interest. In upsampling a random sample is drawn from the observations of the minority class and added to the training set. On the other hand, in downsampling of the majority class, a random subsample of the majority class smaller than the original set is used for training. A more sophisticated method is proposed by Chawla, Bowyer, Hall, and Kegelmeyer (2002) and has shown better classification performance compared to the simple approaches. It is called Synthetic Minority Over-sampling Technique (SMOTE) and is implemented in R (Chawla et al., 2002). This approach uses a combination of downsampling the majority class and the synthetic creation of new instances of the minority class. The model takes the difference between an observation and its nearest neighbours, multiplies this difference by a random number between 0 and 1 and adds it to the feature vector. This will create a new random observation between two observed points. For nominal variables, the feature is given the value that occurred in the majority of the nearest neighbours.

3.2.5 Evaluation measures

To evaluate the predictions made by a classifier, several evaluation metrics can be used. To compute them, the model is applied to a new set of previously unseen data and predictions are obtained. These are compared with a reference or true label. This results in a contingency table or confusion matrix as depicted in Figure 3.1.

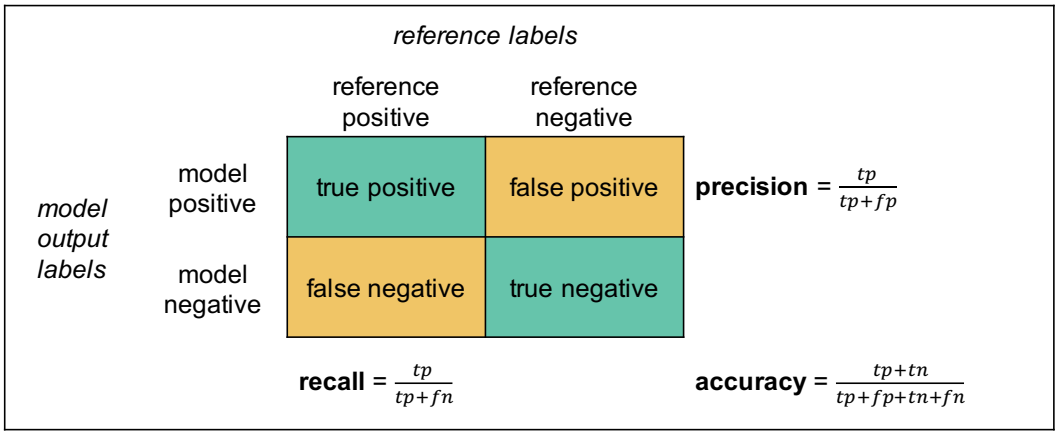


Figure 3.1: Confusion matrix. Contingency table of possible outcomes with the metrics recall, precision and accuracy explained. Abbreviations: tp = true positive, fp = false positive, fn = false negative, tn = true negative. Adapted from Jurafsky and Martin (2019), p. 66.

See (3.5) for an overview of evaluation metrics. Each of the measures captures a different aspect of the classification result of a model. The accuracy is the proportion of correctly labelled observations. This measure has to be interpreted carefully when the classes are

unbalanced, because a classifier that predicts always the majority class will still have a very high accuracy. Recall and precision instead concentrate on the true positives. Recall (or sensitivity) is the proportion of correctly identified positives out of all truly positive observations. Precision is the proportion of correctly labelled positives out of all observations labelled as positive. Specificity instead depicts the proportion of correctly identified negatives of all truly negative observations. In a multi-class classification setting, the confusion matrix for one class considers this class as the positive class and all other classes as the negative classes. This leads to an inflated measure of specificity because the number of true negatives increase.

Metrics that combine two other measures give a better picture of the classification performance. Balanced Accuracy for example is the mean of sensitivity and specificity, and the F1-measure is the harmonic mean of precision and recall. The F1-measure is a conservative measure, because the harmonic mean favours the lower of its two values, and it is an appropriate measure also with unbalanced classes (Jurafsky & Martin, 2019).

In a multi-class setting (with exactly one label per observation), two ways of computing the evaluation metrics are possible. In macro-averaging, the model is tested in a "one versus all" way for every single class and the performances are averaged over classes. This way, all classes are weighted equally. In micro-averaging, the predictions for all classes are collected in one contingency table and the metrics are computed from this table. Here more frequent classes will dominate the results.

The metrics used in this work are listed in (3.5). For multi-class predictions the macro-averages are retrieved and reported.

$$\begin{aligned}
\text{Accuracy} &= \frac{tp + tn}{tp + fn + fp + tn} \\
\text{Balanced Accuracy} &= (\text{sensitivity} + \text{specificity})/2 \\
\text{Sensitivity } (= \text{Recall}) &= \frac{tp}{tp + fn} \\
\text{Specificity} &= \frac{tn}{fp + tn} \\
\text{Precision} &= \frac{tp}{tp + fp} \\
\text{F1} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}
\end{aligned} \tag{3.5}$$

3.3 Implementation and procedure of data analysis

In this section a detailed overview of the steps of the analyses will be provided. All analyses were conducted with R, version 3.6.1 by R Core Team (2019). Data handling was mainly done using the `data.table` format (Dowle & Srinivasan, 2019) and functions of the `tidyverse` (Wickham et al., 2019) and plots were created using the package `ggplot2` (Wickham, 2016). For the imputation of diagnosis codes with the deep learning approach by Biessmann et al. (2018) the package `DataWig` was used, which is provided in Python. This part of the computation was therefore conducted with Python version 3.7.4 (Van Rossum & Drake, 2009). A complete list of the packages used in this work can be found in Table A.4 in Appendix A. The procedure is additionally depicted in the flow chart in Figure 3.2.

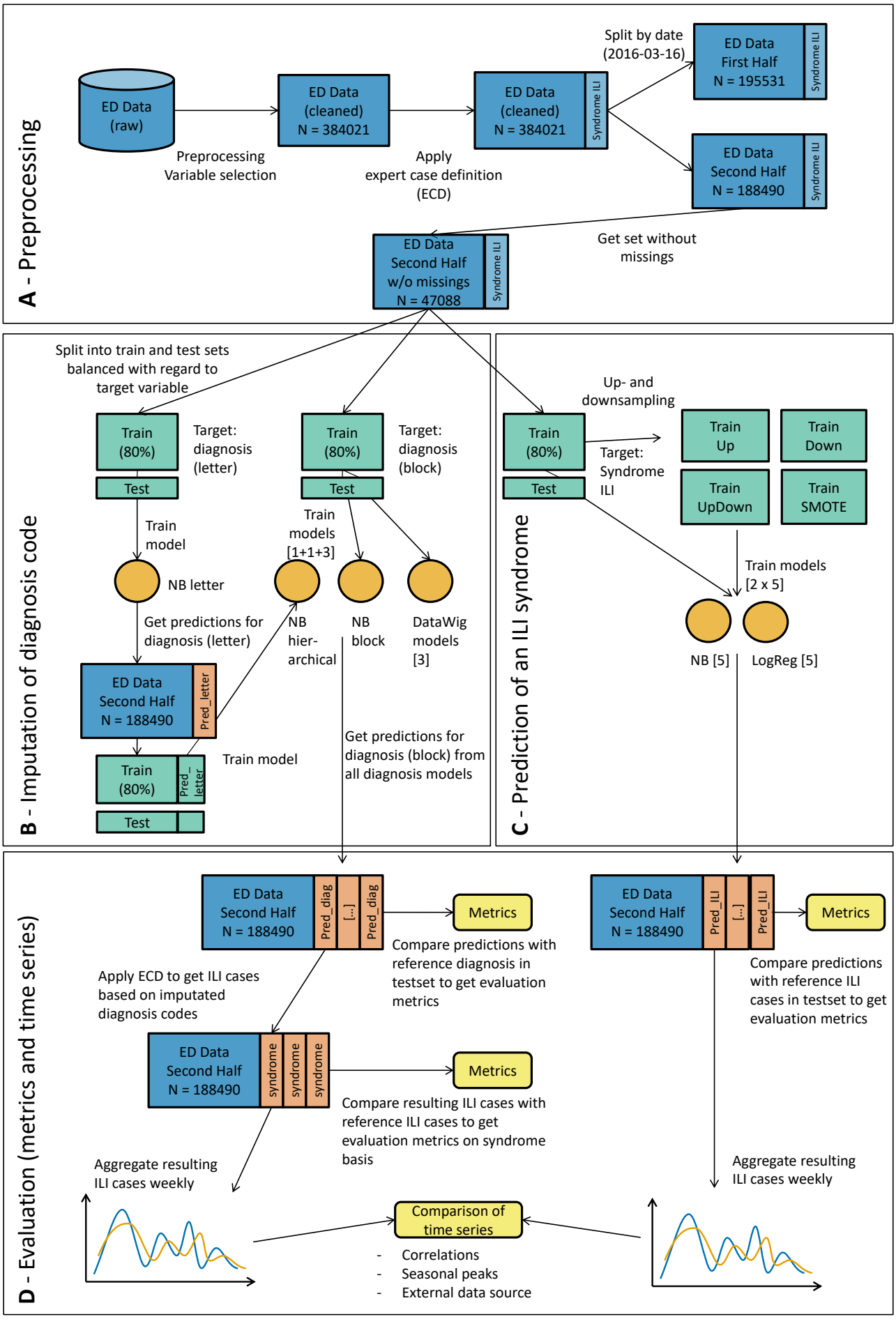


Figure 3.2: Procedure flow chart. For detailed descriptions of the procedures A - Preprocessing: see Section 2.2; B - Imputation of diagnosis codes: see Section 3.3.2; C - Prediction of an ILI syndrome: see Section 3.3.3; D - Evaluation (metrics and time series): see Section 3.3.4. Furthermore, for a description of the train-test-set partitioning see Section 3.3.1 and for a description of the up- and downsampling methods see Section 3.3.3.

Original dataset				
Dataset				N
(a)	Whole set	(2012 - 2019)	with missings	384021
(b)	First half	(2012 - 2016)	with missings	195531
(c)	First half	(2012 - 2016)	without missings	18013
(d)	Second half	(2016 - 2019)	with missings	188490
(e)	Second half	(2016 - 2019)	without missings	47088

Train- and testsets			
Dataset		N_{train}	N_{test}
(f)	Trainset based on diagnosis (block)	37976	9112
(g)	Trainset based on diagnosis (letter)	37680	9408
(h)	Trainset based on syndrome ILI	37671	9417

Table 3.1: Dataset sizes. Sizes of the original datasets (upper part of the table) and the train- and testsets based on different target variables (lower part of the table). "With missings" means that any of the variables can include missing values. "Without missings" means that observations are excluded that have missing values in any of the predictor variables (except for diagnosis).

3.3.1 Train- and test datasets

After the preprocessing described in Section 2.2, the data was split in two parts, with the first half containing cases from 2012 until 2016-03-16, and the second half containing cases from this date up to 2019. After this, cases with missing values in either of the predictor variables were removed, leaving a set of complete cases (A - Preprocessing in Figure 3.2). The sizes of all datasets used in this work are listed in Table 3.1. All models were trained on the second half of the whole dataset containing only cases with no missing values (as described in Section 2.3, *(e)* in Table 3.1). This is due to the fact that some models are not able to deal with missing values in the dataset during training. In a further analysis one might impute these missing values as well, prior to training the models. The reduced dataset was divided into a train and a test set with a ratio of 80:20 (green boxes in parts B and C of Figure 3.2). Because the classification tasks use different target variables (diagnosis (block), diagnosis (letter) or syndrome ILI), the split leads to different train and test sets for each task. The partitioning was implemented using the `createDataPartition()` function of the `caret` package in R, which balances the train and test sets according to the target variable (Kuhn, 2020). The remaining 20% of the dataset is used for testing the models and obtaining evaluation metrics to estimate their behaviour on new datasets. The resulting train and test set sizes are depicted in the lower part of Table 3.1.

The predictor variables for all tasks in this work are age (metric), gender (factor), time (factor), weekday (factor), month (factor), department (factor), referral (factor), complaint group (factor), complaint value (factor), body temperature (median, metric), systolic blood pressure (median, metric), oxygen saturation (median, metric), heart rate (median, metric) and respiratory rate (median, metric). All variables are described in detail in Section 2.2.

3.3.2 Imputation of missing diagnosis codes

The first goal of this thesis is to explore two approaches to impute missing values in the variable diagnosis (*B - Imputation of diagnosis code* in Figure 3.2). The 3-digit diagnosis (block) was

used as the target variable, resulting in 1135 possible classes. The flat naive Bayes classifier and the DataWig models were trained on dataset *(f)* in Table 3.1.

As a baseline approach, a multi-class naive Bayes classifier was used. The model was built using the `naiveBayes()` function in the package `e1071` by Meyer, Dimitriadou, Hornik, Weingessel, and Leisch (2019). To prevent probabilities of exactly 0, a Laplace smoothing factor of $\alpha = 1$ was applied.

As a second approach, the DataWig package by Biessmann et al. (2018) as described in Section 3.1.3 was used. The package requires the user to specify the type of the target and input variables and select the encoding and featurizer methods. Three different encoding methods and featurizers for the complaint variables were compared. In the first version, the complaint variables were considered as being categorical, thus a one hot encoding was used by the model. In the second and third version, the complaint variables were considered as being sequential. The second approach uses a bag-of-words featurizer and the third a LSTM featurizer.

Additionally, a hierarchical approach was designed to factor in the structure of ICD-10 codes, following the work of Perotte et al. (2014). In a first step, a multi-class naive Bayes classifier was trained to predict only the 1-digit diagnosis (letter), i.e. the higher level chapter of the diagnosis (left part in box *B* of Figure 3.2). This model was used to get a prediction of the diagnosis (letter) for every visit in the dataset. Subsequently, a second model was trained to predict the 3-digit diagnosis (block) based on all variables plus the already predicted chapter as an additional input variable.

To compare the models, they were applied to a test set of formerly unseen data and the predicted labels were compared with the reference labels. These counts of correctly and falsely classified cases are used to obtain macro-averaged metrics using the `confusionMatrix()` function in the `caret` package in R (Kuhn, 2020). This is depicted in Figure 3.2 in part *D - Evaluation* in the upper left part.

Because the overall goal is to use the imputed diagnosis codes for further analysis, e.g. detecting the ILI syndrome, the models from this section were evaluated with regard to this task as well. The models were used to get predictions of diagnosis codes for every case in the whole dataset (*(d)* in Table 3.1). Then the expert case definition (see Section 2.4) was applied to the predicted diagnosis codes and the resulting ILI cases were compared to those that result from applying the expert rules to the original diagnosis codes. The results were compared with the same evaluation metrics as before, with regard to the models' abilities to enable the detection of ILI cases (Part *D - Evaluation* of Figure 3.2, below the upper left box).

3.3.3 Classification of an ILI syndrome

The second goal of this thesis is to train a model that can automatically predict the ILI syndrome for a given visit (*C - Prediction of an ILI syndrome* in Figure 3.2). The prediction should be based on all available information in the data except for the diagnosis variable, to make the model independent of the existence of this variable.

To get a labelled training dataset, the expert rule-based case definition (Section 2.4, which is based on diagnosis (block), complaints and temperature) was applied to the dataset (see *A - Preprocessing* in Figure 3.2). In the training set (*(h)* in Table 3.1) the diagnosis (block) variable was removed and the models were trained on the remaining variables. The idea is that the remaining variables will contain enough information on whether a case has ILI or not, thus resulting in a model that can predict the ILI syndrome without relying on the diagnosis variable.

Two models were compared: a binary naive Bayes classifier and a logistic regression. The

binary naive Bayes model was trained using the `naiveBayes()` function in the package `e1071` and a laplace smoothing factor of $\alpha = 1$ was used (Meyer et al., 2019). The logistic regression was performed with the `train()` function of the package `caret` by Kuhn (2020), with `method = "glm", family = binomial(link = "logit")`.

Resampling Because the two classes ILI versus not ILI are highly unbalanced in the dataset (3.34% ILI cases), a classifier might not be able to detect the less prevalent class very precisely. To make a classifier more sensitive to the minority class, different resampling methods can be used. In this case, five different methods are compared to each other.

Simple Upsampling: A random sample with replacement of 20000 cases was drawn from the minority class and added to the train set, resulting in a train set with n = 57671 cases of which 36.86% are ILI cases.

Simple Downsampling: A random sample without replacement of 20000 cases was drawn from the majority class and used together with the existing ILI cases to form a train set with n = 21259 cases of which 5.92% are ILI cases.

Up- and Downsampling: The random sample of 20000 cases from the minority class was combined with the random sample of 20000 cases from the majority class, resulting in a train set with n = 40000 cases of which 50% are ILI cases.

SMOTE: The `SMOTE()` function in package `DMwR` by Torgo (2010) was used to simultaneously create a synthetic sample of new minority cases and delete some of the majority class cases in order to get a more balanced dataset. This resulted in a train set with n = 57914 cases of which 41.30% are ILI cases.

Original dataset: The original train set has a size of n = 37671 cases (*(h)* in Table 3.1) with 3.34% being ILI cases.

Five individual models for each classification method are compared to each other (green boxes in part *C* of Figure 3.2). The models were applied to the test set and the predicted labels are compared to the reference labels to obtain evaluation metrics (upper right part of box *D* in Figure 3.2).

Variable importance One goal of this work was to make the decisions made by the models comprehensible for users like epidemiologists and thus increase acceptance for this approach. For the two models used here, logistic regression and naive Bayes, the model output already gives a good insight on how the models come to a conclusion. For the logistic regression model, the z-values of the regression coefficients can be interpreted as the importance of each variable for the prediction of the model. The z-value is the regression coefficient divided by its standard error. If this value is significantly different from 0, it indicates that the corresponding predictor is relevant for the prediction. The direction of the deviation (positive or negative) corresponds to the direction of the predictors influence. The 20 most important (= largest z-values) predictors from all five logistic regression models were extracted and their corresponding z-values reported.

The naive Bayes models contain the posterior conditional probabilities for each predictor, which can be used to reconstruct the models' decisions. Since there is no simple method of visualizing them at this point of time (for an implementation in Orange see Možina, Demšar, Kattan, and Zupan (2004)), reporting them would go beyond the scope of this results section. For interested readers the results can be provided nevertheless.

3.3.4 Evaluation of weekly aggregated cases

In the third step, the two approaches (*B* and *C* in Figure 3.2) to deal with missing diagnosis codes are evaluated with respect to the task of predicting the ILI syndrome. The resulting weekly aggregated ILI cases are compared (bottom part of box *D - Evaluation (metrics and time series)* in Figure 3.2). The dataset used for this evaluation step is the second half (*(d)* in Table 3.1) with all originally available data. Because there are no diagnosis codes available for all cases and because the ILI case definition used in this study can not be considered "gold-standard", no real ground truth is available for comparisons. Instead it is assumed that the ILI cases that result from applying the rule-based case definition on the dataset (even if diagnosis codes are missing) represent the ground truth and these cases are considered as "internal reference" in the further analysis. The detected ILI cases are aggregated on a weekly basis, resulting in time series of frequencies for all approaches and models. As a metric for comparison the Pearson correlation between the time series is reported.

Selection of the best models In a final step, those models have to be selected that are best suited to find ILI cases when the diagnosis code is missing. Selection criteria derived from the introductory and methods parts are the following: A good model

1. should be able to find as many ILI cases as the original rule-based definition as possible (as is reflected by the sensitivity (recall) in an evaluation on syndrome basis);
2. while still having a reasonable precision (as reflected in the aggregated F1-measure);
3. should preferably predict more ILI cases instead of less than the original approach (because the internal reference ILI cases might not be a perfect ground truth, it is tolerable if the models result in more ILI cases);
4. should follow the same seasonal pattern as the original ILI cases;
5. should be positively correlated with the original ILI cases at a weekly aggregated level;
6. should be easy to interpret and offer a comprehensible decision making process;
7. and should be positively correlated with the external data source (hospital SARI cases).

The different models are evaluated with regard to the first five of these criteria and the best models for each approach (diagnosis imputation, syndrome prediction) are selected. The final evaluation steps are performed only on these models.

Evaluation of false classifications The models will most likely have false predictions, meaning that they will predict a case to have ILI even if the expert rule says it does not (false positive) or predict a case to not have ILI even if the expert rule says it does (false negative). These false classifications might help to evaluate the appropriateness of the models. Therefore they are examined on a medical level to estimate the degree of the model being wrong. The diagnosis codes from the false positives (fp) and false negatives (fn) are categorized by a medical expert regarding their likeliness to be an ILI case. If the diagnosis originally belonged to the expert definition, the case was labelled with "ILI yes". If the diagnosis of a fp or a fn is not in the original expert rule, but might been given in an ILI case as well, it was labelled "ILI possible". Only if it can be ruled out that this diagnosis is not related to ILI in any way, the case is labelled "ILI no".[7]The false positives and false negatives of the best models

are evaluated on this first medical level, and secondly the distribution of the diagnosis codes themselves into the ICD chapters are examined.

[7] An example for "ILI yes" is the code J10 (Influenza with pneumonia, seasonal influenza virus identified). A case with this code is very likely to have ILI. An example code for "ILI possible" is J20 (Acute bronchitis). This code does not belong to the original ILI definition in this study, but is closely related to ILI and may be given to a case with similar symptoms. An example for "ILI no" is the code J15 (Bacterial pneumonia, not elsewhere classified), which does involve the respiratory system, but was ruled out to be related to influenza by the medical expert. Another example is A08 (Viral and other specified intestinal infections), which involves an infectious disease as well, but is not related to influenza.

4 Results and evaluation

This section is divided into three parts. First, the results of imputing the diagnosis codes are shown. This is followed by the results of the classification of the ILI syndrome. Last, all models are evaluated on a time series basis and the best models are selected.

4.1 Imputation of missing diagnosis codes

In this section, the results for imputing the ICD-10 diagnosis codes will be presented. They were obtained using a flat naive Bayes classifier, the DataWig package and a hierarchical naive Bayes model. All evaluation metrics can be found in Table 4.1.

4.1.1 Evaluation of imputed diagnosis codes

The naive Bayes classifier predicting the diagnosis (block) yielded a F1-measure of 28.75%, 22.82% precision and 7.17% recall. The balanced accuracy was 53.53%. Of the three DataWig models using different preprocessing and feature extraction methods, the one using a bag-of-words approach yielded the best results. It had a F1-measure of 34.37%, 32.32% precision and 11.07% recall. The balanced accuracy was 55.46%. Both other DataWig approaches had comparable results and all of them were better in predicting the correct ICD codes than the naive Bayes model (block). The hierarchical naive Bayes model showed an improved performance compared to the flat naive Bayes model (30.33% F1-measure, 25.68% precision, 8.71% recall and 54.29% balanced accuracy). The auxiliary naive Bayes model predicting the diagnosis (letter), that is then used by hierarchical naive Bayes model, had a F1-measure of 38.80%, 25.68% precision, 42.12% recall and 69.71% balanced accuracy.

Model	Balanced Accuracy[a]	Sensitivity (Recall)[a]	Specificity[a]	Precision[a]	F1[a]
Naive Bayes					
Block	0.5353	0.0717	0.9986	0.2282	0.2875
Hierarchical	**0.5429**	**0.0871**	0.9986	**0.2568**	**0.3033**
Letter[b]	*0.6971*	*0.4212*	*0.9730*	*0.3934*	*0.3870*
DataWig					
Categorical	0.5483	0.0980	0.9985	0.2981	0.3360
Bag of words	**0.5546**	**0.1107**	**0.9986**	**0.3232**	**0.3437**
LSTM	0.5490	0.0994	0.9985	0.3108	0.3093

[a] Macro-averaged metrics (all classes weighted equally).
[b] Auxiliary model for NB Hierarchical.

Table 4.1: Evaluation metrics for the models imputing the ICD diagnosis codes. Macro-averaged metrics are shown for the flat and the hierarchical naive Bayes classifier, the auxiliary naive Bayes classifier predicting the diagnosis (letter), and the DataWig models. The highest value of each metric is highlighted for each group of models.

4.1.2 Evaluation of the syndrome cases based on imputed diagnosis codes

To evaluate the use of the imputation models for finding ILI cases, imputed diagnosis codes were used to obtain ILI cases by the expert case definition. These cases were compared to

Model	Balanced Accuracy	Sensitivity (Recall)	Specificity	Precision	F1	$tp + fp$[a]
Naive Bayes						
Block	0.6108	0.2256	**0.9961**	**0.6380**	0.3333	1805
Hierarchical	**0.6162**	**0.2371**	0.9953	0.6038	**0.3404**	2025
DataWig						
Categorical	0.6592	0.3333	0.9850	0.4045	0.3655	4399
Bag of Words	0.6699	0.3518	**0.9880**	**0.4714**	**0.4029**	3905
LSTM	**0.6730**	**0.3595**	0.9865	0.4483	0.3990	4333

[a] $tp + fp$ = total amount of predicted cases. Dataset size: n = 188490; $tp + fn$ = 4704 (total amount of reference ILI cases).

Table 4.2: Evaluation metrics resulting for diagnosis models on syndrome level. The ILI cases are compared to those that result when the expert rule is applied to the imputed diagnosis codes by each imputation model. The highest value of each metric is highlighted for each group of models.

the internal reference cases, which result from the original diagnoses. This was done for the whole dataset of the second half, including cases with missing values (dataset *(d)* in Table 3.1). The models showed varying results as can be seen in Table 4.2. F1-measure was best for the DataWig bag-of-words approach (40.29%), recall and balanced accuracy were best in the DataWig LSTM model (35.95% and 67.30% respectively) and precision was best in the flat naive Bayes model (63.80%). Apart from precision and specificity, the hierarchical naive Bayes model yielded better results than the flat naive Bayes classifier. All models detected less ILI cases compared to the reference in the second half of the time frame, where the original (rule-based) amount of ILI cases is 4704. Both naive Bayes models resulted in less than half that many cases (1805 and 2025) and the DataWig models resulted in 3905 (bag of words) to 4399 (categorical) cases.

4.2 Classification of an ILI syndrome

A naive Bayes classifier and a logistic regression model were compared in their ability to predict an ILI syndrome that is based on the expert case definition without relying on the diagnosis code. Several up- and downsampling methods were used. See Table 4.3 for the evaluation metrics of all syndrome models.

4.2.1 Evaluation of the syndrome predictions

Of the naive Bayes models, recall and balanced accuracy were best in the model using up- and downsampling (NBUpDown), with 79.62% recall and 82.02% balanced accuracy. Precision and F1-measure were best in the normal naive Bayes model (25.17% precision and 32.97% F1-measure). In overall, the two models NB and NBDown had similar metrics and the three models NBUp, NBUpDown and NBSMOTE had comparable metrics, diverging in whether recall and balanced accuracy are better or precision and F1-measure. All models classified more cases as ILI as were originally present in the dataset (314 cases). The naive Bayes model (without resampling) and NBDown predicted approximately the same amount of ILI cases and two times the originally present ILI cases (601 and 731), but NBDown had a higher recall (53.82% vs. 47.77%). NBUp, NBUpDown and NBSMOTE predicted about five times as many cases as originally present (1485, 1719 and 1512 versus 314).

Model	Balanced Accuracy	Sensitivity (Recall)	Specificity	Precision	F1	$tp + fp$[a]
Naive Bayes						
NB	0.7144	0.4777	**0.9510**	**0.2517**	**0.3297**	601
NBDown	0.7385	0.5382	0.9387	0.2325	0.3247	731
NBUp	0.8107	0.7548	0.8665	0.1632	0.2684	1485
NBUpDown	**0.8202**	**0.7962**	0.8442	0.1499	0.2523	1719
NBSMOTE	0.8003	0.7389	0.8618	0.1557	0.2572	1512
Logistic Regression						
logReg	0.5994	0.2038	**0.9949**	**0.5818**	0.3019	105
logRegDown	0.6625	0.3376	0.9874	0.4796	**0.3963**	216
logRegUp	0.8336	0.7707	0.8964	0.2042	0.3229	1209
logRegUpDown	**0.8387**	**0.8312**	0.8462	0.1571	0.2643	1708
logRegSMOTE	0.8142	0.7452	0.8831	0.1803	0.2903	1271

[a] $tp + fp$ = total amount of predicted cases. Testset size: n = 9417.

Table 4.3: Evaluation metrics for syndrome models. The prediction results for the naive Bayes and the logistic regression models are shown, with the normal model and four resampling methods for each approach. In the last column the total amount of predicted ILI cases for each model is shown, which can be compared to the original number of ILI cases in the test set ($n_{ILI} = 314$). The highest value of each metric is highlighted for each group of classifiers.

For the logistic regression model, the results differ. The two models logReg and logRegDown classify about a third and half of the cases as ILI compared to the originally present cases (105 and 216 vs. 314). Precision is highest for the normal logistic regression model (58.18%) and F1-measure is best for the logistic regression using downsampling (39.63%). LogRegUp, logRegUpDown and logRegSMOTE classify more cases as ILI than originally present (1209, 1708, 1271 vs. 314). Of these three models, logRegUp has the highest F1-measure (32.29%). The three models have higher recall values than the first two models, with logRegUpDown having the highest recall (83.12%) and balanced accuracy (83.87%).

4.2.2 Variable importance of the logistic regression models

To get an understanding of which variables were important for the models to predict whether a visit has ILI or not, the absolute z-value of the logistic regression models are reported. For the logistic regression models, the 20 most important variables for each model are shown in Figure 4.1. Note that categorical variables were transformed to dummy variables by the models and therefore each category is depicted as a variable in this plot. All of the models have the body temperature (median) as their most important variable. Of the complaint groups, the most frequent ones are malaise in adults (*Unwohlsein bei Erwachsenen*), headache (*Kopfschmerzen*), malaise in children (*Unwohlsein bei Kindern*), general indicators (*Generelle Indikatoren*), respiratory distress in children (*Atemnot bei Kindern*), respiratory distress in adults (*Atemnot bei Erwachsenen*), conspiciuous behaviour (*Auffälliges Verhalten*) and sore throat (*Halsschmerzen*). The months February, March and April are among the most important variables. All of these variables were also significant on a 0.001-level, indicating that the regression coefficients are statistically significant different from zero. The most important variables are approximately the same for the first four models in Figure 4.1, only for logRegSMOTE they have a different order. For logRegUp and logRegUpDown more variables are statistically significant different from zero than for logReg and logRegDown.

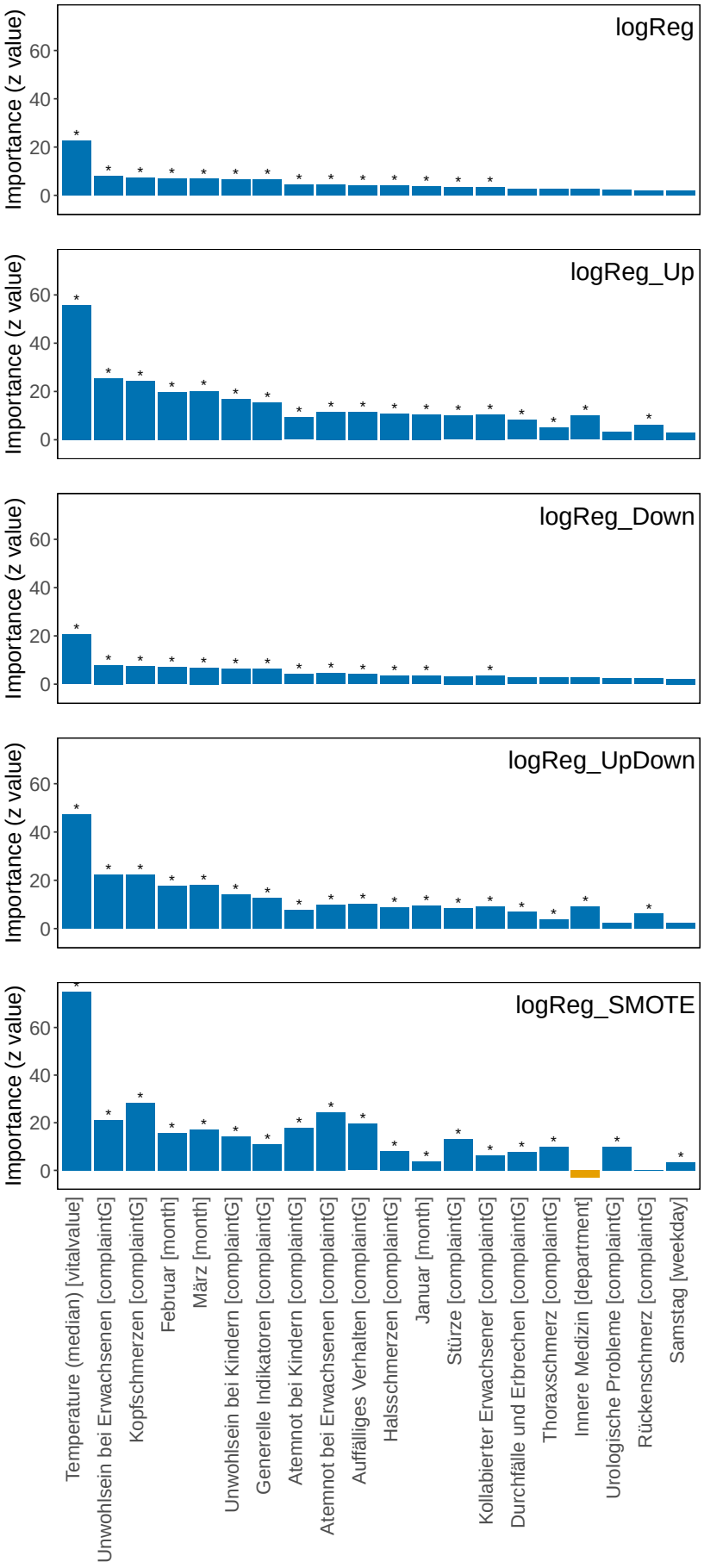
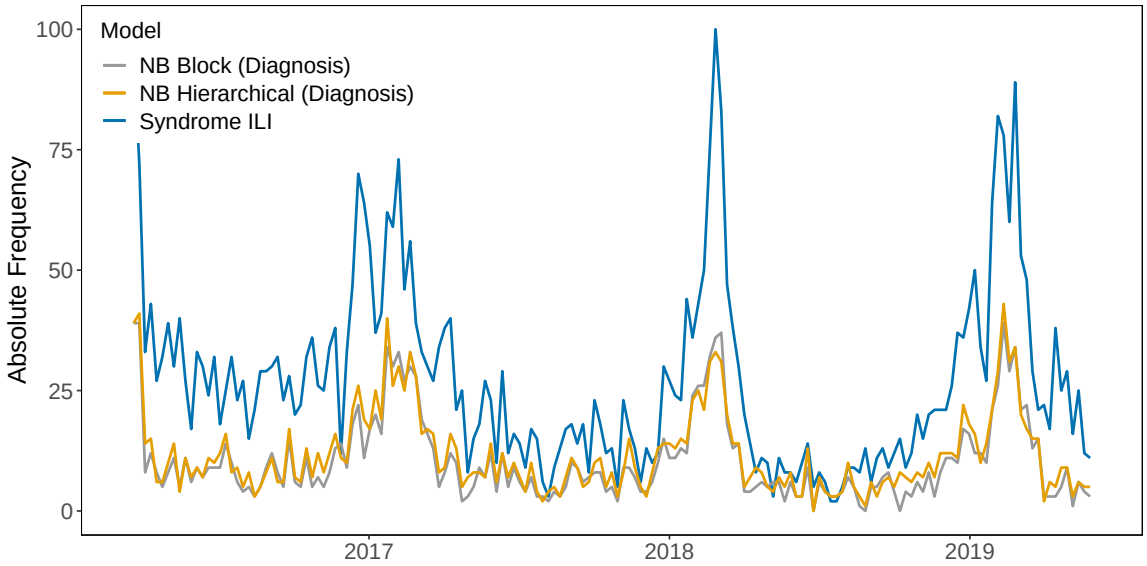


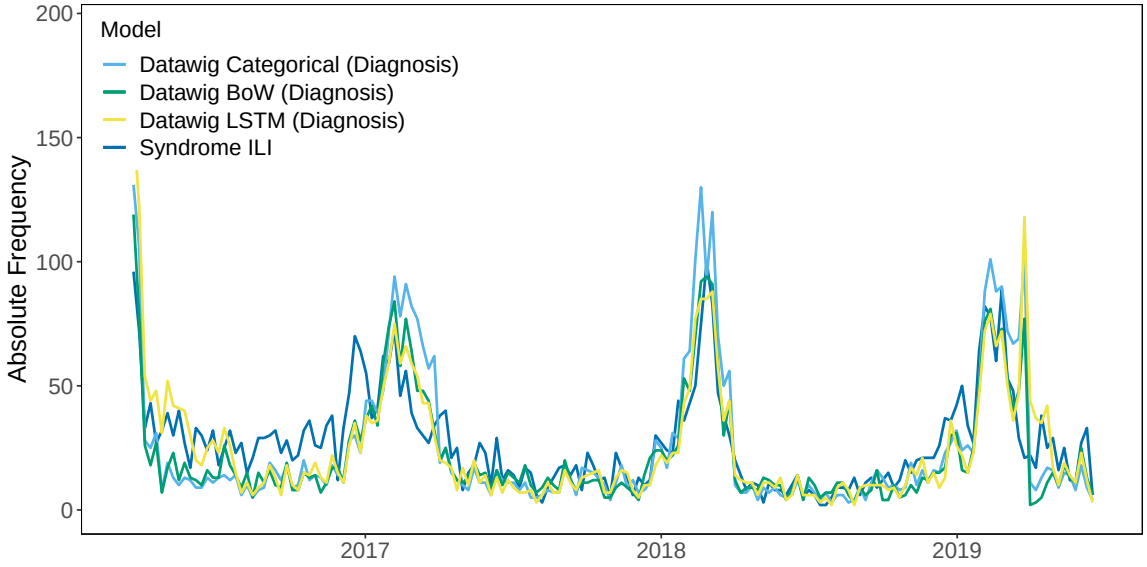
Figure 4.1: Variable importance of logistic regression models. The absolute z-value for the 20 most important variables in each model is shown. * $p < 0.001$ (for reasons of readability).

4.3 Evaluation of weekly aggregated cases

For syndromic surveillance it is important to monitor the change of ILI cases over time. This is why a further evaluation was implemented on the resulting ILI case time series. To obtain the time series for the imputed diagnosis codes, the expert case definition was applied to the imputed codes and the resulting ILI cases were aggregated weekly. For the syndrome prediction models, the predicted ILI cases were aggregated weekly as well. Both were compared to the reference time series, that is obtained by applying the expert case definition to the originally present diagnosis codes (Syndrome ILI).



(a) Weekly aggregated cases of naive Bayes imputation models



(b) Weekly aggregated cases of DataWig imputation models

Figure 4.2: Weekly aggregated ILI cases resulting of the imputation models. The ILI cases are shown that result from applying the expert rule on the imputed diagnosis codes by the imputation models, compared to the reference (dark blue line). **(a)** naive Bayes (normal vs. hierarchical approach) and **(b)** DataWig models (three different encoders). The absolute amount of cases is shown, for the second half of the dataset.

4.3.1 Imputed diagnosis codes

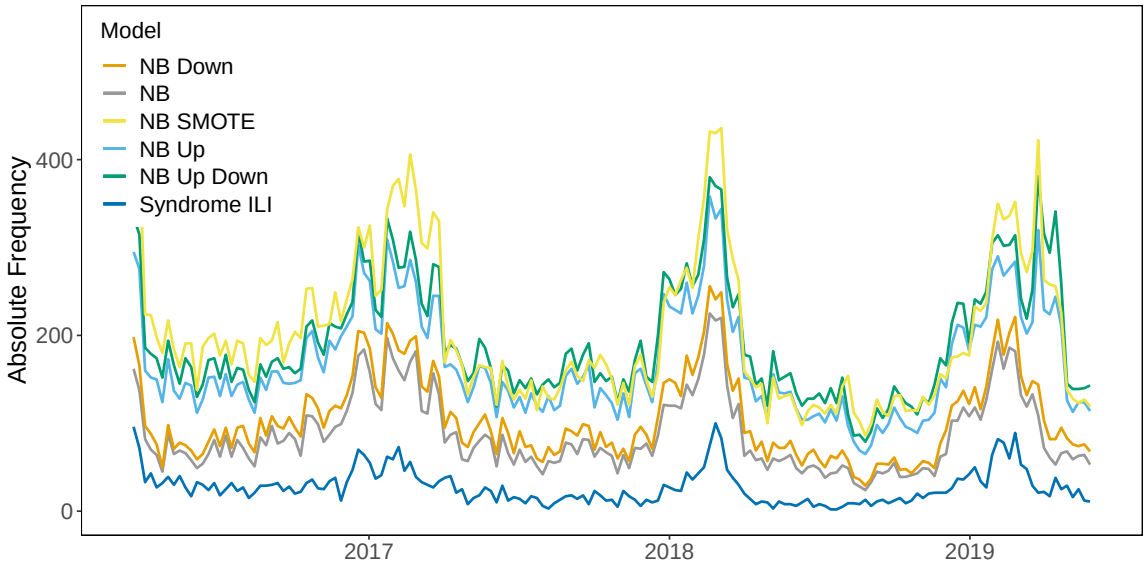
Figure 4.2a shows the weekly ILI cases resulting from the imputation models using the flat versus the hierarchical naive Bayes classifier versus the reference (dark blue line). It can be seen that both resulting time series follow approximately the same line and are both below the reference for most of the time. The weekly aggregated cases have a correlation of $r = 0.873$ ($p < 0.001$, NB Block) and $r = 0.865$ ($p < 0.001$, NB Hierarchical) with the reference. They show the same seasonal pattern with peaks in January, February and March.

In Figure 4.2b the weekly cases resulting from the DataWig imputation models are shown. The time series resemble the reference and follow the same seasonal pattern. The correlations are $r = 0.807$ ($p < 0.001$, DataWig Categorical), $r = 0.839$ ($p < 0.001$, DataWig BoW) and $r = 0.797$ ($p < 0.001$, DataWig LSTM). There is no clear difference between the models, but the BoW-Model has the highest correlation with the reference.

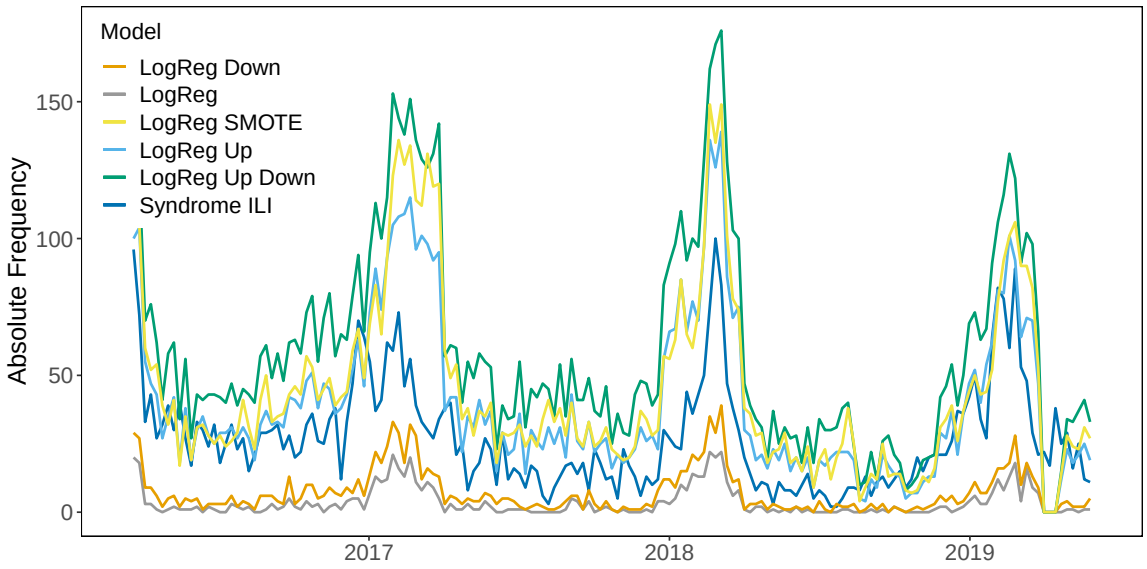
4.3.2 Syndrome models

Figure 4.3a shows the weekly aggregated ILI cases predicted by the naive Bayes models from the second part of the analyses. As we saw in Section 4.2, all naive Bayes models predicted more cases compared to the expert definition. This is clearly visible in the time series graphs. The normal naive Bayes model (NB) and the one using downsampling (NBDown) predicted around the same number of cases and the models NBUp, NBUpDown and NBSMOTE predicted around the same number. All models follow the same seasonal pattern as the reference. The weekly cases from the normal naive Bayes model correlate highest with the reference ($r = 0.866$, $p < 0.001$), followed by the model using downsampling ($r = 0.861$, $p < 0.001$). NBUp, NBUpDown and NBSMOTE have the following correlations with the reference: $r = 0.831$ ($p < 0.001$), $r = 0.799$ ($p < 0.001$) and $r = 0.856$ ($p < 0.001$).

Figure 4.3b shows the weekly aggregated ILI cases resulting from the logistic regression models predicting ILI. The two models logReg (normal) and logRegDown predicted less cases compared to the expert definition, and both time series are clearly below the dark blue line for the reference. The other three models predicted more cases as ILI and are exceeding the reference line in most of the times, with bigger abbrevations in the first two years and less in 2019. The weekly cases from the logRegDown model correlate most with the reference with $r = 0.812$ ($p < 0.001$), followed by logRegUp with $r = 0.793$ ($p < 0.001$). The normal logReg model, logRegUpDown and logRegSMOTE have the following correlations with the reference cases: $r = 0.789$ ($p < 0.001$), $r = 0.789$ ($p < 0.001$) and $r = 0.769$ ($p < 0.001$).



(a) Weekly aggregated cases of naive Bayes syndrome classifiers



(b) Weekly aggregated cases of logistic regression syndrome classifiers

Figure 4.3: Weekly aggregated ILI cases predicted by the syndrome classifiers. The predicted amount of cases are compared to the reference (dark blue line). **(a)** Naive Bayes models and **(b)** the logistic regression models, each with four different up- and downsampling methods. The absolute amount of cases is shown, for the second half of the dataset.

4.3.3 Selection of the best models

At this point a selection of the models was obtained. The goal was to find the models from each approach that are best suited to find ILI cases, when the information about the diagnosis code is missing. For an overview of the selection criteria see Section 3.3.4. The models are selected based on criteria 1. - 5. and then the selected models are further evaluated.

From the first approach, where imputation models were used to fill up missing diagnosis codes, the hierarchical naive Bayes classifier showed better recall and F1-measures than the flat naive Bayes classifier both when predicting the ICD codes and when evaluated on syndrome level. Of the DataWig models, the LSTM approach had a higher recall on syndrome level, but the bag-of-words approach had a better F1-measure and performed best on diagnosis-level.

Even though the hierarchical NB classifier had the highest correlation with the reference time series, it predicted less ILI cases than present in the reference (2025 vs. 4704 in the testset). The DataWig BoW model predicted only slightly less ILI cases than the reference (3905 vs. 4704) and had a higher recall (35.18% vs. 23.71% on syndrome level), therefore this model might be better suited to detect possible ILI cases. These two approaches (hierarchical naive Bayes and DataWig BoW) were chosen as the best models from the imputation approach, even though both resulted in less ILI cases than originally present.

The syndrome classification models were described in detail as well. From the naive Bayes models, the normal model (NB) and NBDown did not predict as many ILI cases as the last three models with 601 and 731 versus 314 original cases (and 1485, 1719 and 1512 for the last three models). Both models followed the same seasonal pattern as the reference (see Figure 4.3a), with correlations of $r = 0.866$ and $r = 0.861$ respectively. Furthermore both models had approximately the same F1-measure (32.97% for NB and 32.47% for NBDown). Because NBDown had a higher recall (53.82%) than normal NB (47.77%), it was chosen as the best naive Bayes model.

Of the logistic regression models, the first two (normal and logRegDown) resulted in less ILI cases than in the reference, whereas the last three resulted in more ILI cases. Of these three models, logRegUp has the highest correlation with the reference ($r = 0.793$). It has the highest F1-measure (32.29%), while still having reasonable recall and balanced accuracy measures and predicting as many ILI cases as logRegUpDown and logRegSMOTE do. It is therefore chosen as the best logistic regression model.

4.3.4 Further evaluation of the selected models

The models selected in the last section were then further evaluated. They were used in the first half of the dataset, where no diagnosis codes are available at all. Furthermore, they were compared to another data source, the ICOSARI data, to validate them externally. A comparison with the scenario of no available diagnosis codes at all was done to show the advantage of the models over the ICD code-based case definition. As a last step, the false classifications of the models were evaluated.

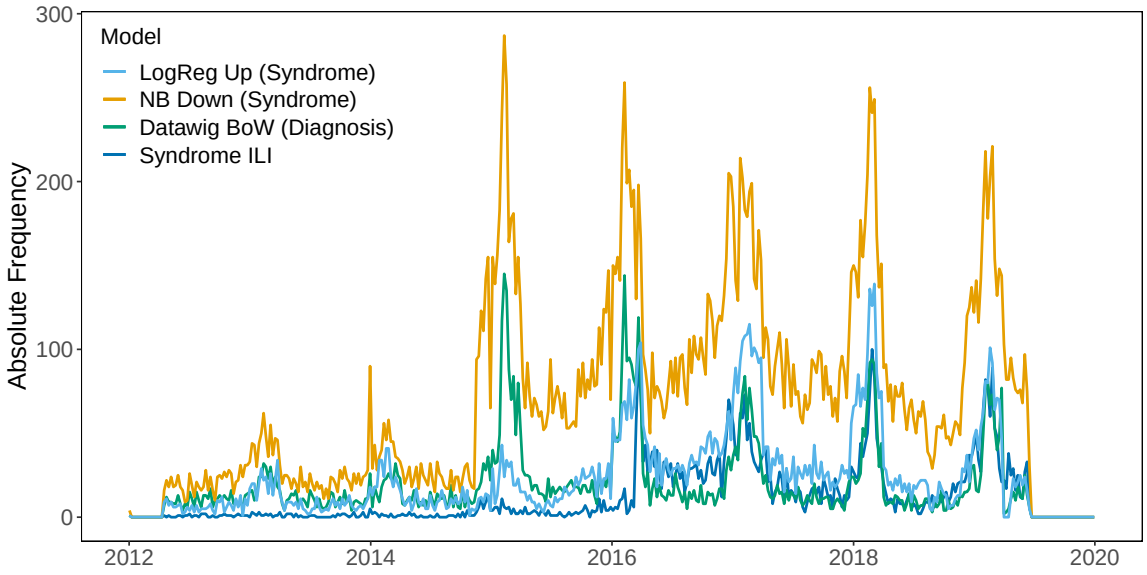


Figure 4.4: Weekly aggregated ILI cases of the best models for the whole dataset. Predictions were obtained for the first and the second half of the dataset and the resulting absolute number of ILI cases are shown. Notice that for the expert rule (dark blue line) almost no cases are detected in the first half of the dataset, because the diagnosis code is missing for all visits.

Performance in first half of the dataset The first half of the dataset poses a use case with no available diagnosis codes. Here the newly developed models might help to detect ILI cases that otherwise would be unable to find. Figure 4.4 shows the resulting ILI cases from the best models for the whole time. It can be seen that the blue line, indicating the internal reference (using the rule expert definition on the existing diagnosis codes), results in almost no ILI cases up to March 2016. This is of course due to the fact, that no diagnosis codes are available in this period of time. For the influenza season in 2015 and 2016, the new models enable to estimate the course of the ILI waves. In the first years (2012 - 2014), the models are not able to extract the same peaks of ILI cases. This might be due to missing data in the rest of the dataset as well, especially in the complaint variables, which are most informative for the classifications. Still the models find more ILI cases without relying on the diagnosis code than the expert definition.

Comparison to an external data source In Section 2.5 it was shown that the ILI cases derived by the case definition in this study can be compared to the severe acute respiratory infection cases of the ICOSARI data base. The weekly aggregated cases from this data source can be used as an external validating reference for the models, especially in the first half of the data, where no internal reference can be derived.

Table 4.4 depicts the correlations of the best models with the internal reference (Syndrome ILI) and the external reference (ICOSARI data) for the second half of the dataset. The syndrome model NBDown correlates highest with the external data source, ICOSARI ($r = 0.867$, $p < 0.001$). The other models have a high correlation with the external data as well, and the internal reference has the lowest correlations with the ICOSARI data. A plot with the best models compared to the external ICOSARI data can be seen in Figure 4.5. The models NBHierarchical (Diagnosis) and NBDown (Syndrome) have the highest correlations with the internal reference ($r = 0.865$, $p < 0.001$ and $r = 0.861$, $p < 0.001$).

	Syn. ILI	NBDown	LogRegUp	NBH.	DataWigBoW	ICOSARI
Syndrome ILI	1					
NBDown	0.861***	1				
LogRegUp	0.793***		1			
NBHierarchical	0.865***			1		
DataWigBoW	0.839***				1	
ICOSARI	0.786***	0.867***	0.821***	0.783***	0.831***	1

*** $p < 0.001$

Table 4.4: Correlations of the internal reference ILI cases with the best models and the ICOSARI data. The resulting ILI cases from the best models were compared with the internal reference (Syndrome ILI) for the second half of the dataset.

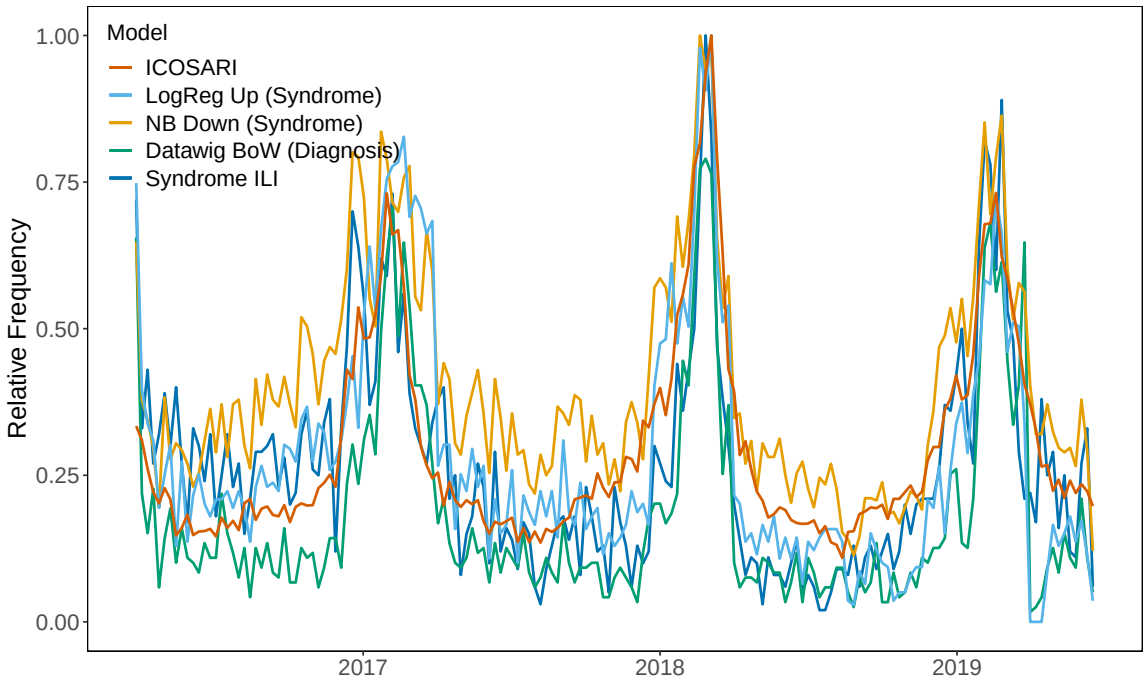


Figure 4.5: Weekly aggregated ILI cases of the best models compared to ICOSARI data. The relative amount scaled to the maximum of each time series is shown, for the second half of the dataset.

While using the expert case definition on the first half of the data set produces almost no ILI cases and results in a correlation of $r = 0.561$ ($p < 0.001$) with the ICOSARI data (2014-01-01 to 2016-03-15), all models explored in this thesis are able to detect at least some ILI cases in the first half. Correlations between the model ILI cases and ICOSARI data for the first half (2014-01-01 to 2016-03-15) are: $r = 0.825$ ($p < 0.001$) for NBHierarchical, $r = 0.815$ ($p < 0.001$) for NBDown, $r = 0.809$ ($p < 0.001$) for DataWigBoW and $r = 0.639$ ($p < 0.001$) for logRegUp. It can therefore be shown that the models follow the external data source with almost the same precision in the first half of the dataset as they do in the second half. This can also be seen as an indicator that the models perform well on unseen data. A plot with the best models over the whole time frame together with the external data can be seen in Figure A.2 in Appendix A.

Comparison to no diagnosis codes existing at all To get another impression of how useful the models explored in this thesis will prove, they are compared to the case when no diagnosis codes are available at all. This is depicted in Figure 4.6. The grey line at the bottom indicates this worst case, when the case definition based on ICD codes leads to a low amount of ILI cases and thus will not be helpful for syndromic surveillance. The benefit of the different models becomes obvious.

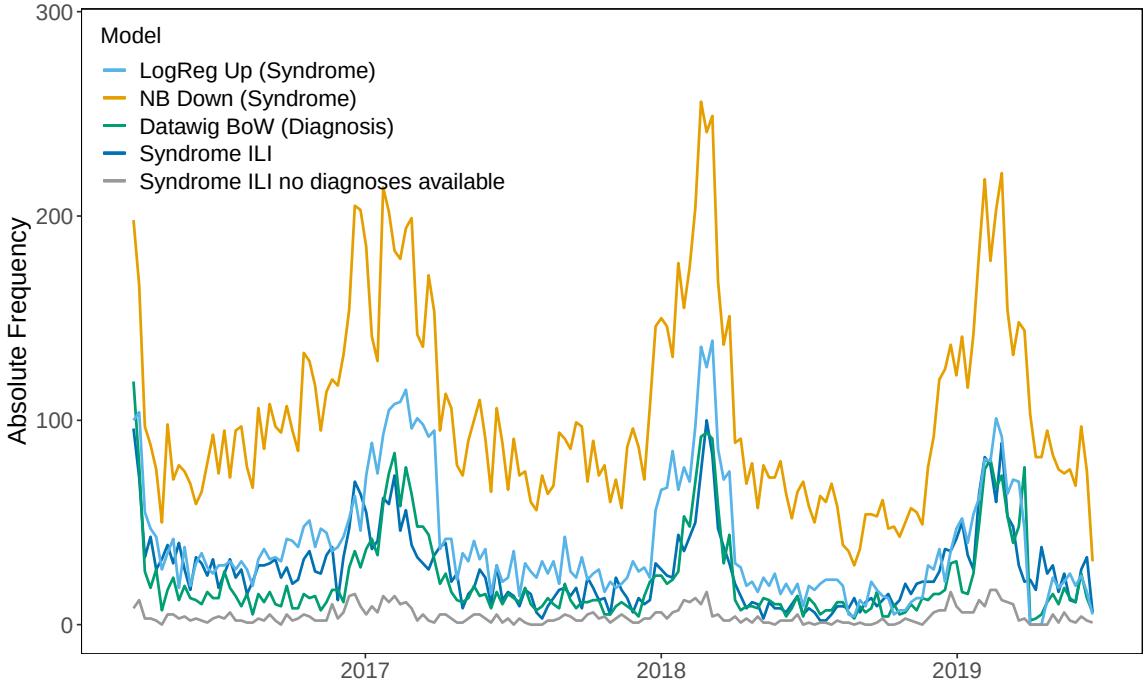


Figure 4.6: Weekly aggregated ILI cases of the best models compared to the scenario with no diagnosis codes available. The grey line shows the resulting ILI cases for the expert rule if no diagnosis codes are available. The absolute number of cases is shown, for the second half of the dataset.

Evaluation of false classifications In this section, the selected models are explored in detail with regard to their false classifications. For the false positives and false negatives, the most frequent diagnosis codes are examined. Additionally, the diagnosis codes are grouped into categories according to how likely they are related to ILI (see Section 3.3.3). This was done to enable a quick sanity check and assess the quality of the models.

We know from Tables 4.2 and 4.3 that both imputation models (hierarchical NB and DataWigBoW) resulted in less than the originally present ILI cases, whereas the classification models logRegUp and NBDown resulted in a higher number of ILI cases. The false positives of the four selected models labelled with ILI yes, possible or no, are depicted in Table 4.5. In all of the models, the majority of the cases can be clearly labelled as non-ILI cases.[8] The NBDown classification model and the DataWigBoW model have a higher number of possible ILI cases in their false positive classification than the other two models (35.23% and 36.87% versus 19.28% and 21.86%). All models have a very low number of actual ILI cases in their false positives, which is to be expected. To get a more detailed insight into the false classifications, the diagnosis codes of these false positives were looked at with regard to the letter or chapter of the diagnosis (see Table 4.6).

[8]It has to be kept in mind that this still relies on the first given diagnosis code that is not necessarily the only relevant diagnosis.

Most of the false positives have diagnosis codes from the chapters J (Diseases of the respiratory system), R (Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified), I (Diseases of the circulatory system), N (Diseases of the genitourinary system), A and B (Certain infectious and parasitic diseases). Chapter J includes most of the actual influenza-related diagnoses and chapter R includes various symptoms instead of diagnoses, like cough or headache. The most frequent diagnosis code of the false positives is J44 (Chronic obstructive pulmonary disease with acute lower respiratory infection Excl.: with influenza (J09-J11)). It can be assumed that this diagnosis implies very similar symptoms compared to influenza-like illness and is therefore either predicted as ILI by the classification models or might get another diagnosis code from the chapter J assigned by the imputation models. The second most frequent diagnosis of false positives is J15 (Bacterial pneumonia), where the same explanation applies.

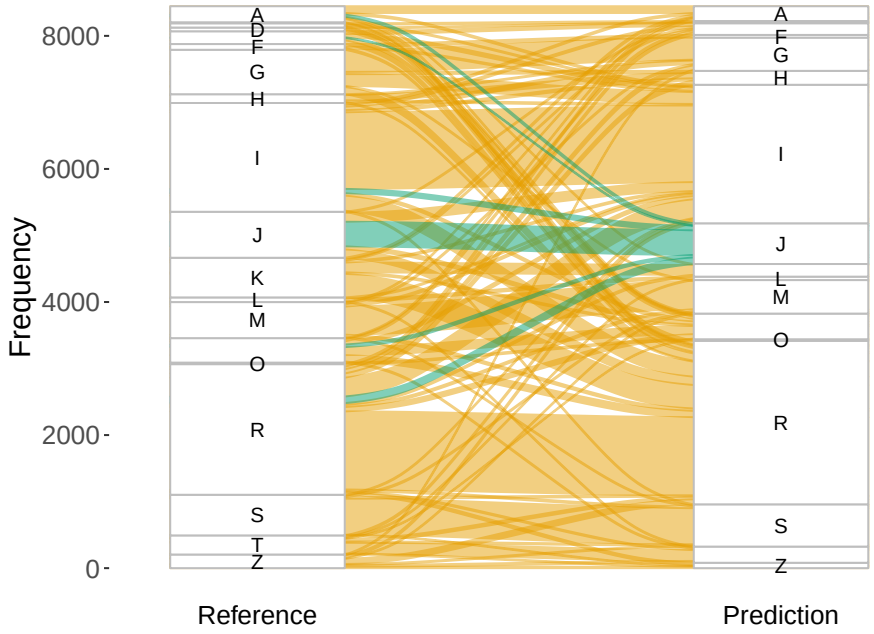


Figure 4.7: Alluvial plot for the predicted diagnoses of DataWigBoW. Those cases are shown in green that got a diagnosis (block) of the chapter J by the DataWigBoW model. On the left side, the original chapter of these diagnoses is shown.

These findings indicate that overall the classification models are able to predict diseases that are medically related to each other, but might not be able to distinguish with the precision needed. Especially for the diagnosis prediction (but similarly for the syndrome prediction) this might be due to low information density in the other available variables. Symptoms coded in MTS are probably the most informative predictors, but still only have 51x191 different categories and are probably not able to represent the same amount of information that a physician uses for making a decision.

The third most frequent diagnosis of false positives is N39 (Other disorders of urinary system, including Urinary tract infection). This is a somewhat surprising finding, but the misclassification might be due to similar symptoms. When examining the most frequent symptoms in these cases, 60.25% of them had symptoms that were related to ILI as well or were unspecific symptoms.[9]

[9]For the false positives of the syndrome model logRegUp, that had N39 as their first diagnosis (n = 317), the most frequent complaint group and complaint value combinations were: malaise in adults & adult

	logRegUp		NBDown		DataWigBoW		NBHierarchical	
Category	N	%	N	%	N	%	N	%
ILI no	3699	79.41	7375	63.58	1139	62.58	554	77.16
ILI possible	898	19.28	4087	35.23	671	36.87	157	21.86
ILI yes	61	1.31	138	1.19	10	0.55	7	0.98
Sum	4658	100.00	11600	100.00	1820	100.00	718	100.00

Table 4.5: False positives of the best models, by category. The cases falsely predicted as being ILI by the best models are evaluated on a medical level. The absolute number and percentages of cases definitely being ILI (ILI yes), possibly being ILI (ILI possible) or definitely not being ILI (ILI no) are shown.

The alluvial plot in Figure 4.7 provides an insight in the predictions of one of the imputation models. It can be seen, that for visits that get a J diagnosis as the prediction, the original diagnoses usually are J, R, I, A, E and N diagnoses. These were already identified as the most frequent diagnoses chapters of false positives, therefore providing an explanation why they were labelled as ILI based on the imputed diagnosis codes but not based on the original diagnosis codes.

The same evaluation was done for the false negatives, i.e. the cases classified as non-ILI, that are originally ILI cases. The categorisation of the false negatives into ILI yes, possible or no is depicted in Table 4.7 and the categorisation into diagnosis (block) can be found in Table 4.8. All models had classified approximately 5% of cases as non-ILI that were actually no ILI cases (regarding the diagnosis code). Around 95% of all false negatives were verified as being actual ILI cases. In Table 4.8 it can be seen that the the diagnosis codes of the false negative cases are just a few codes, with the most common codes of cases that originally should have been classified as ILI being J18, J06, J10 and R50.

temperature $> 38.5°C$ (n = 85), malaise in adults & recent problem (< 7 days) (n = 28), malaise in adults & rapid onset (n = 27), general indicators & adult temperature $> 38.5°C$ (n = 26) and respiratory distress in adults & low oxygen saturation (n = 25).

Diagnosis (letter)	logRegUp		NBDown		DataWigBoW		NBHierarchical	
	N	%	N	%	N	%	N	%
A	372	7.99	1024	8.83	158	8.68	80	11.14
B	113	2.43	609	5.25	145	7.97	45	6.27
C	72	1.55	45	0.39	20	1.10	13	1.81
D	88	1.89	87	0.75	20	1.10	13	1.81
E	192	4.12	110	0.95	40	2.20	24	3.34
F	14	0.30	87	0.75	7	0.38	3	0.42
G	71	1.52	134	1.16	17	0.93	15	2.09
H	26	0.56	684	5.90	74	4.07	5	0.70
I	565	12.13	316	2.72	61	3.35	47	6.55
J	1266	27.18	4440	38.28	691	37.97	155	21.59
K	193	4.14	377	3.25	60	3.30	27	3.76
L	28	0.60	159	1.37	12	0.66	5	0.70
M	58	1.25	79	0.68	20	1.10	15	2.09
N	478	10.26	534	4.60	165	9.07	108	15.04
O	8	0.17	30	0.26	15	0.82	19	2.65
P	5	0.11	209	1.80	4	0.22	1	0.14
Q	1	0.02	89	0.77	1	0.05	0	0.00
R	894	19.19	1568	13.52	173	9.51	84	11.70
S	35	0.75	234	2.02	9	0.49	12	1.67
T	48	1.03	211	1.82	18	0.99	12	1.67
X	0	0.00	1	0.01	0	0.00	0	0.00
Z	131	2.81	573	4.94	110	6.04	35	4.88
Sum	4658	100.00	11600	100.00	1820	100.00	718	100.00

Table 4.6: False positives of the best models, by diagnosis (letter). The cases falsely predicted as being ILI by the best models are evaluated on a second medical level. The absolute number and percentages of the diagnosis codes (letter) of the false positives are shown.

Category	logRegUp		NBDown		DataWig06bow		NBHierarchical	
	N	%	N	%	N	%	N	%
ILI no	20	5.62	72	5.34	136	4.55	189	5.37
ILI possible	0	0	1	0.07	0	0	0	0
ILI yes	336	94.38	1276	94.59	2854	95.45	3332	94.63
Sum	356	100.00	1349	100.00	2990	100.00	3521	100.00

Table 4.7: False negatives of the best models, by category. The cases falsely predicted as not being ILI by the best models are evaluated on a medical level. The absolute number and percentages of cases definitely being ILI (ILI yes), possibly being ILI (ILI possible) or definitely not being ILI (ILI no) are shown.

Diagnosis (block)	logRegUp		NBDown		DataWig06bow		NBHierarchical	
	N	%	N	%	N	%	N	%
J06[a]	65	18.26	541	40.10	1665	55.67	1921	54.56
J09[b]	1	0.28	3	0.22	2	0.07	4	0.11
J10[a]	61	17.14	188	13.94	264	8.83	328	9.32
J11[b]	19	5.34	66	4.89	128	4.28	178	5.06
J12[b]	0	0.00	2	0.15	6	0.20	7	0.20
J18[a]	169	47.47	372	27.58	559	18.70	630	17.89
J20[c]	0	0.00	1	0.07	0	0.00	0	0.00
J22[a]	9	2.53	26	1.93	44	1.47	58	1.65
J44[b]	0	0.00	1	0.07	0	0.00	0	0.00
R05[a]	0	0.00	7	0.52	24	0.80	40	1.14
R50[a]	32	8.99	142	10.53	298	9.97	355	10.08
Sum	356	100.00	1349	100.00	2990	100.00	3521	100.00

[a] ILI yes
[b] ILI no
[c] ILI possible

Table 4.8: False negatives of the best models, by diagnosis (block). The cases falsely predicted as not being ILI by the best models are evaluated on a second medical level. The absolute number and percentages of the diagnosis codes (block) of the false positives are shown.

5 Discussion

The aim of this thesis was to investigate two approaches for dealing with missing values in the diagnosis variable in an emergency department dataset, to subsequently be able to detect cases that might have an influenza-like illness syndrome. To achieve this, in the first approach, three methods were compared that impute the missing ICD-10 diagnosis codes from all other variables. A flat multi-class naive Bayes classifier was compared to a hierarchical naive Bayes classifier and the imputation package DataWig. In the second approach, a binary classification model was trained to predict the syndrome ILI for a given visit, without using the diagnosis code as input. A naive Bayes classifier and a logistic regression model were compared and several resampling methods were used to tackle the high imbalancedness of the training dataset. All models were evaluated with regard to their ability to detect the ILI cases as accurately as possible and in a last step compared on a weekly aggregated time series level. The best models from each approach were selected and further examined.

In the following, the results of the imputation and classification approaches as well as the evaluation of the models are briefly summarized and discussed. The limitations of this analysis are considered and further developments are suggested.

5.1 Critical discussion of the results

Imputation of diagnosis codes The classification and imputation of the diagnosis codes worked reasonably well and showed that predicting clinical diagnoses in an emergency department dataset from all information linked to a patient stay can help enhancing the dataset for subsequent analyses. The DataWig imputation models performed slightly better than the naive Bayes classifiers, both on model- and syndrome-level. Only on time series-level the naive Bayes models had minimal higher correlations with the reference time series. Of the DataWig models, the bag-of-words approach performed best regarding F1-measure and correlation with the reference time series. The DataWig models that interpreted the complaint variable as sequential were better able to predict the diagnosis code than when interpreting complaint as categorical. This might be due to the format of the complaints, that consist of several words or a short description of the symptom. When regarded as representations in a feature space some complaints might have smaller or larger distances to each other, caused by a similar wording. This can be taken into account in the prediction process, but would not be recognised if taken as categorical features. If in the future more data sources are added to the system (e.g. patient records), this approach has the key advantage that even sequential variables can be used as input to impute the diagnosis codes without the need of further preprocessing steps.

On the other hand, the naive Bayes models differ only slightly from the DataWig models in terms of F1-measure and recall, indicating that even this rather simple model is able to keep up with the more complex machine learning approach. This is in line with other findings of the performance of the naive Bayes classifier in difficult tasks compared to other methods and emphasizes its predictive performance when most of the input variables are categorical (Rish, 2001). The NB classifier was furthermore shown to perform well when predicting medical diagnoses (Al-Aidaroos, Bakar, & Othman, 2012; Pakhomov et al., 2006; Scheurwegs et al., 2017). Compared to the neural network-based method in DataWig, naive Bayes offers an easily understandable model and decision making process. Furthermore it requires way less computational power and is much faster. If additional sequential data is available in the future, a naive Bayes approach can be used for classifying this data as well, since it performs well for text classification tasks (Mitchell, 1997).

The hierarchical naive Bayes classifier resulted in better predictions than the flat classifier,

indicating that taking into account the hierarchical nature of the ICD-10 diagnosis codes in the prediction process can enhance the performance, as suggested by Perotte et al. (2014) and Zhang (2008). The procedure explored in this work represents a rather simplistic way of implementing the hierarchy into the model. It was used to get a first impression of whether this might improve the performance. In a future analysis one could implement a procedure similar to the one proposed by Perotte et al. (2014), with one classifier predicting the first hierarchical level (diagnosis (letter)) and then 23 classifiers on the second level, one for each letter. Another idea could be to take into account the hierarchy in the evaluation metrics, for example by inserting a factor for the correct prediction of the first level (diagnosis (letter)). Nevertheless, the flat and the hierarchical naive Bayes classifiers from this work showed comparable results to those of Perotte et al. (2014), where the flat (SVM) classifier had an F1-measure of 27.6% (28.7% in this work) and the hierarchical 39.5% (30.33% in this work).

Comparing the imputation models on syndrome- and time series-level showed that the reference ILI cases could be detected reasonably well by the expert definition using the imputed diagnosis codes. All time series have the same seasonal peaks around February and March of each year. This also emphasizes the use of ED data for the syndromic surveillance of influenza-like illness, as suggested by Zheng et al. (2007) and Bourgeois, Olson, Brownstein, McAdam, and Mandl (2006).

Regarding the evaluation metrics reported in this work, it has to be taken into consideration that for a multi-class setting the metrics can be computed in different ways. In this work, the macro-averaged metrics were computed, which weight each class the same (versus micro-averaging, where larger classes are weighted more compared to smaller classes). Some metrics can therefore be affected by imbalanced class sizes, if the model performs differently for small versus large classes. If a model has better performance for larger classes, a macro-average is a more conservative metric. If it performs better on smaller classes, a macro-average might be an overestimation. Therefore, the macro-averages suggest a more conservative picture of the overall performance. In this work, when predicting diagnosis codes, it is considered more relevant that the model performs well for the large classes which are the diagnosis codes mostly given and present in an emergency department. When looking into the relationship of e.g. recall in a single class (diagnosis (block)) compared to this classes' size (see Figure A.3 in Appendix A) for the DataWigBoW model, we can indeed see that recall is better for larger classes ($r = 0.485$, $p < 0.001$). F1-measures on the other hand are not correlated with class frequency ($r = 0.140$, n.s.), confirming that F1-measures are appropriate for unbalanced classes. To get a better picture of the performance of the models without relying too heavily on the smaller classes, one might also have a look at the micro-averaged metrics.

It can be concluded that the imputation models propose a valid approach of filling up missing values in the diagnosis variable and thereby enable subsequent analyses based on the diagnosis codes. The hierarchical approach improved the performance compared to a flat classifier. A health outcome like ILI can be detected with a reasonable accuracy, enabling the assessment of seasonal peaks on a large time frame. However, for predictions on single case level, the models might not be suited that well. In this case, a multi-label approach would be a more appropriate solution, which allows for multiple diagnoses to be possible and with an expert making the final decision. This could be implemented with the current dataset as well, with the advantage that several diagnosis codes are already existing for one visit.

Prediction of an ILI syndrome The second part of this thesis consisted of learning and predicting the ILI syndrome from all available variables except for the diagnosis code. It could be shown that this way of automatically finding a case definition for ILI that is independent of the diagnosis code worked quite well, but the models have to be evaluated cautiously with

regard to their decision making process. For the logistic regression models, the relevant predictor variables were identified. The body temperature was the most important predictor in all of the models, together with the symptoms mailaise in adults (*Unwohlsein bei Erwachsenen*) and headache (*Kopfschmerzen*). These predictors can be considered relevant for the ILI syndrome and were confirmed by a physician and epidemiologist. The twelve most important variables (see Section 4.2.2) can be considered relevant for the ILI syndrome and were approved by epidemiologists. After these variables, some models had complaints like falls (*Stürze*), collapsed adult (*Kollabierter Erwachsener*) and diarrhoea and vomiting (*Durchfälle und Erbrechen*) as significant predictors. These can not be linked medically to the ILI syndrome, but probably occur often in an emergency department setting. Furthermore, the months February, March and January are among the twelve most important predictors for all models. Whereas this is theoretically sound, the inclusion of the months into the models in the first place is questionable. The models might label a case with ILI-like symptoms as "ILI" only if it appears within these months. Or it might falsely classify another infection as ILI because it appears within these months. Because the goal of the model is to detect ILI based on the symptoms (and other health related data), it might not be suitable to include this variable in the first place. The models might need to be trained without this variable to confirm their ability to detect relevant ILI cases.

The different resampling methods for balancing the minority versus majority class in the training set improved the performance of the models compared to the basic ones based on the original training set. Simple downsampling resulted in models with higher sensitivity than and almost the same specificity as the normal models. Upsampling, combined up- and downsampling and the SMOTE algorithm resulted in higher sensitivity but smaller specificity of the models compared to the normal ones. Interestingly, the SMOTE algorithm did not improve the prediction of ILI more than did the simple up- and downsampling methods. This might be caused by the fact that while the SMOTE algorithm does include an upsampling method for nominal data as well it was originally designed for numeric data only (Torgo, 2010). With all of these resampling methods it has to be kept in mind that the sampling was only repeated once. To improve the stability of the results one could draw multiple random samples for up- and downsampling and average the resulting metrics across them, but also vary the amount of sampled data.

Regarding the time series of weekly aggregated ILI cases by the classification models, it can be seen that the predicted cases by both model types follow the seasonal peaks around January to March. Furthermore, the NB models detect much more ILI cases compared to logReg or the reference ILI cases. This amount could be adjusted by changing the threshold at which a case is classified as ILI. To assess the relationship between sensitivity and specificity for each model, ROC curves could be examined as well. In ROC curves the sensitivity is plotted against 1 - specificity (Fawcett, 2006). This can help to find the optimal threshold in order to obtain the desired sensitivity or specificity. Because logistic regression and naive Bayes allow to get a prediction of a probability, it is possible to set a threshold manually and therefore adjust the model to the desired sensitivity and specificity. Additionally, the priors of the minority class can be changed for the naive Bayes classifier, therefore making it more sensitive to this class (Chawla et al., 2002).

It has to be kept in mind that the naive Bayes classifiers will always predict the outcome for any given observation, even if it has missing values in one of the variables. The logistic regression models on the other hand will only make a prediction if the observation has valid values in all of the variables of the model. Logistic regression would need an imputation of the other missing values before applying the model, but one could also use the model on an incomplete dataset. The results from the time series evaluation suggested that even with only

considering the complete cases, the logistic regression models (e.g. logRegUp) can give an appropriate impression of the existing ILI cases.

In general, both classification model types are also dependent on the labels in the training set and how they were determined. In this case, a more sensitive, unspecific ILI case definition was used to label the cases, to obtain a larger number of positive training examples. This leads to an inclusion of visits with more unspecific symptoms, making the model learn to include these cases in the prediction. This might be one of the reasons why that many false positives of the logRegUp and NBDown models were considered non-ILI cases. It might therefore be interesting to use a more specific set of rules for the labelling of the reference ILI cases, resulting in only few but very clear ILI cases. The training set would need to be upsampled again to obtain enough positive training examples. Because several additional datasets of EDs exist in the project, this kind of analysis might be feasible to conduct, combined with a downsampling of the non-ILI cases.

An advantage of the logistic regression model is that interactions can be included. This might help to model the relationship better, for example by including an interaction term with temperature and the symptom variables. Non-linear relationships were not considered in this analysis and might be worth to take into consideration as well.

In conclusion, it can be seen that the ILI syndrome can be predicted with the two classification models relying on the variables from the emergency department setting and therefore proposing an alternative case definition that is independent of the existence of diagnosis codes. The most important source of information are symptoms, as reported by the complaint group (MTS category). Compared to the first approach, imputation of diagnosis codes (and especially the DataWig models), the syndrome models provide a better explainability and can be extended to other syndromes as well. Still, a new model has to be trained for each syndrome, whereas the imputation models only need to be executed once.

Performance on the first half and comparison with external data When using both approaches, imputing diagnoses and predicting the ILI syndrome, to obtain ILI cases in the first half of the dataset, where no diagnosis codes are available at all, the advantage of these two approaches becomes very clear. Relying solely on the expert definition based on ICD-10 diagnosis codes, no analyses would be possible for this period of time. Both approaches were able to detect some ILI cases to allow for subsequent analyses. It was also shown that in this period of time the models have a reasonable correlation with the external data source ($r = 0.825$ for NBHierarchical with ICOSARI).

It was surprising to find that the data from this study resembled an external data source that well. First of all, this study only includes data from one single emergency department, not many across Germany. The ICOSARI data on the other hand joins data from about 80 hospitals all over Germany. Second, the case definitions used in the ICOSARI data differs from the ILI case definition used in this study. Still the cases seem to be comparable, indicating that the ILI definition in this study captures also less severe influenza cases and cases with acute respiratory infections, which in turn resemble those cases included by ICOSARI (hospitalised/more severe acute respiratory infections). The ICOSARI data therefore served as a useful source of validation and with including more emergency departments in the analysis, the comparison can be enhanced.

Evaluation of false classifications The evaluation of false classifications of all models and the predictors of the logistic regression models revealed that the models learned to predict a syndrome that resembles the ILI syndrome from the expert case definition. Nevertheless, the models included many cases that were not originally labelled as ILI and the evaluation showed

that this is mostly due to very similar symptoms or diagnosis codes. In a future application, the models should be either trained with more specific labels or the threshold should be reduced to include only cases with more severe symptoms. Regarding the false negatives, a more detailed evaluation would bring insights in why they were not found to be ILI cases in the first step of the classification. For the diagnosis imputation models this might be due to the models predicting a non-ILI-related diagnosis code for a case that originally had an ILI-related ICD code. For the ILI prediction models it would be interesting to see what characteristics the false negatives have in comparison to the true positives. And last, the evaluation of false classifications could also be extended to complaints or variables like temperature.

5.2 Limitations of this work

This work has several limitations that will be discussed in this section. First of all, with selecting only the first diagnosis code from all available diagnoses for a visit, a pre-selection and reduction of information is done. Even though several diagnoses might be true for one visit, only one was selected. This can lead to cases where the symptoms might not be clearly related to the diagnosis code. A way to deal with this would be a multi-label approach, where in training and prediction multiple diagnosis codes are possible (see Larkey and Croft (1995) and Pakhomov et al. (2006) for suggested approaches). Even though in this particular setting predicting more than one diagnosis is not feasible, considering the second and maybe third likely diagnosis for the evaluation of the prediction model might improve its performance (as suggested by Perotte et al. (2014)). At the same time, the diagnosis was reduced to the 3-digit version, therefore further diminishing the amount of information. This problem might be solved by a multi-label approach as well, or by a more sophisticated hierarchical approach.

A drawback of the syndrome prediction approach is that it is still dependent on the data being labelled manually before the analysis to have a reference to train the model. An a priori case definition has to be made by epidemiologists and the trained model is dependent on this definitions' sensitivity and specificity. It is therefore not a fully automatic or unsupervised approach, but enables application users to adapt the model or original case definition with regard to the results.

Furthermore, as described in Section 2.4, the expert case definition was originally developed using the 4-digit ICD-codes, but was then applied to the 3-digit codes. In this dataset this resulted in only 12 falsely included cases, but this procedure should not be applied to another dataset without assessing the amount of wrongly included cases before.

The training dataset was derived from a reduced dataset containing only visits with no missing values in the relevant variables at all. This might lead to a selection bias in the resulting cases. An analysis of the distribution of the variables for the whole dataset versus the second half with all cases versus the second half with no missing values revealed that the mean of age differs for the last dataset (see Table A.2 in Appendix A). This might indicate that the values are not completely missing at random, but instead more missing values occur in younger patients. Furthermore, the distribution of cases regarding diagnosis (letter), the involved department and the time of admission was different in the reduced dataset compared to the whole set. In this case, this only influences the generalisability of the models to new datasets, because they might be less sensitive at detecting the syndrome in a different population.

It also has to be considered that the vital parameters were aggregated to a median over the observed values in one visit. Different preprocessing methods are possible as well, e.g. using only the maximum value or binning them to predefined categories with regard to severity.

Another limitation lies in the hierarchical structure of the MTS symptoms. Because only

the most severe complaint is reported, other possible symptoms are neglected. Free-text variables could provide more information, but might take more time to be filled out by the medical staff. Still, this kind of variable is not available in the present dataset.

Last, this study uses data from only one emergency department and therefore the results can not easily be generalised. Because of the general data framework that is used in the proposed syndromic surveillance system, the integration and comparison of additional emergency departments is achieved easily and the models from this study can be explored on a broader set of EDs covering all of Germany. It might be of interest how the models differ between the EDs and whether they can be applied on new datasets.

5.3 Conclusion

This study showed that two different approaches are possible for dealing with missing diagnosis codes in an ED setting, when the goal is to monitor an influenza-like illness syndrome. It was possible to predict and impute the missing diagnosis codes from the available ED data. Numeric and categorical variables could be used, with the potential of expanding the model to include sequential data without further preprocessing steps. The dataset is therefore enhanced and a diagnosis code can be predicted for a visit at any time, enabling further analyses and the syndromic surveillance based on diagnosis codes. The predictions made by the models were sufficiently precise to subsequently monitor an ILI syndrome. Introducing a multi-label approach or considering the hierarchical nature of ICD-codes could improve the performance.

Additional to the existing rule-based case definition, a model for a data-driven case definition derived from a given case definition for the ILI syndrome was developed. This flexible case definition is less vulnerable to missing data and is able to predict the syndrome by relying on symptoms, vital parameters and other information available in the dataset. The sensitivity can be adjusted according to the use case, including more or less severe cases. Furthermore, this generic method can be extended to other syndromes as well.

In conclusion, the syndromic surveillance of an ILI syndrome based on the present ED data was enabled, through enhancing the data basis and developing flexible methods to find ILI cases in a realistic dataset with missing diagnoses. While this study was based on data from only one ED, it will be the next step to apply the models to additional datasets and other syndromes to evaluate their generalisability.

References

Aguilera, J., Paget, W., Mosnier, A., Heijnen, M., Uphoff, H., Van der Velden, J., ... Watson, J. (2003). Heterogeneous case definitions used for the surveillance of influenza in europe. *European journal of epidemiology*, *18*(8), 751–754. doi:10.1023/A:1025337616327

Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical data classification with naive bayes approach. *Information Technology Journal*, *11*(9), 1166–1174.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018). Multi-label classification of patient notes: Case study on icd code assignment. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Bawa, Z., Elliot, A. J., Morbey, R. A., Ladhani, S., Cunliffe, N. A., O'Brien, S. J., ... Smith, G. E. (2015). Assessing the likely impact of a rotavirus vaccination program in england: The contribution of syndromic surveillance. *Clinical Infectious Diseases*, *61*(1), 77–85. doi:10.1093/cid/civ264

Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). Deep learning for missing value imputationin tables with non-numerical data. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 2017–2025). CIKM '18. ACM. doi:10.1145/3269206.3272005

Bourgeois, F. T., Olson, K. L., Brownstein, J. S., McAdam, A. J., & Mandl, K. D. (2006). Validation of syndromic surveillance for respiratory infections. *Annals of emergency medicine*, *47*(3), 265–e1. doi:https://doi.org/10.1016/j.annemergmed.2005.11.022

Buda, S., Tolksdorf, K., Schuler, E., Kuhlen, R., & Haas, W. (2017). Establishing an icd-10 code based sari-surveillance in germany–description of the system and first results from five recent influenza seasons. *BMC public health*, *17*(1), 612. doi:10.1186/s12889-017-4515-1

Bundesamt für Justiz. (1988). Sozialgesetzbuch (SGB) Fünftes Buch (V) - Gesetzliche Krankenversicherung - §295 Abrechnung ärztlicher Leistungen. Retrieved March 7, 2020, from https://www.gesetze-im-internet.de/sgb_5/__295.html

Bundesamt für Justiz. (2000). Gesetz zur Verhütung und Bekämpfung von Infektionskrankheiten beim Menschen. Retrieved March 7, 2020, from http://www.gesetze-im-internet.de/ifsg/index.html

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Casalegno, J.-S., Eibach, D., Valette, M., Enouf, V., Daviaud, I., Behillil, S., ... Cohen, J. M., et al. (2017). Performance of influenza case definitions for influenza community surveillance: Based on the french influenza surveillance network grog, 2009-2014. *Eurosurveillance*, *22*(14). doi:10.2807/1560-7917.ES.2017.22.14.30504

Caserio-Schönemann, C., Bousquet, V., Fouillet, A., & Henry, V. (2014). The french syndromic surveillance system sursaud®. *Bull Epidémiol Hebd*, 3–4.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chini, F., Farchi, S., Ciaramella, I., Antoniozzi, T., Rossi, P. G., Camilloni, L., ... Borgia, P. (2009). Road traffic injuries in one local health unit in the lazio region: Results of a surveillance system integrating police and health data. *International journal of health geographics*, *8*(1), 21. doi:10.1186/1476-072X-8-21

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091. doi:10.1016/j.jclinepi.2006.01.014

Dowle, M., & Srinivasan, A. (2019). *Data.table: Extension of 'data.frame'*. R package version 1.12.8. Retrieved March 13, 2020, from https://CRAN.R-project.org/package=data.table

Elliot, A. J., Bone, A., Morbey, R., Hughes, H., Harcourt, S., Smith, S., ... Andrews, N., et al. (2014). Using real-time syndromic surveillance to assess the health impact of the 2013 heatwave in england. *Environmental research*, *135*, 31–36. doi:10.1016/j.envres.2014.08.031

Espino, J. U., Dowling, J., Levander, J., Sutovsky, P., Wagner, M. M., & Copper, G. (2006). Syco: A probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. In *Syndromic surveillance conference, baltimore, maryland*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Garcia, A. J., & Hruschka, E. R. (2005). Naive bayes as an imputation tool for classification problems. In *Fifth international conference on hybrid intelligent systems (his'05)* (3–pp). IEEE. doi:10.1109/ICHIS.2005.78

Henning, K. J. (2004). What is syndromic surveillance. *Morbidity and mortality weekly report*, *53*(Supplement), 7–11.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hughes, H., Morbey, R., Hughes, T., Locker, T., Shannon, T., Carmichael, C., ... McCloskey, B., et al. (2014). Using an emergency department syndromic surveillance system to investigate the impact of extreme cold weather events. *Public Health*, *128*(7), 628–635. doi:10.1016/j.puhe.2014.05.007

Hughes, H., Morbey, R., Hughes, T., Locker, T., Pebody, R., Green, H., ... (2016). Emergency department syndromic surveillance providing early warning of seasonal respiratory activity in england. *Epidemiology & Infection*, *144*(5), 1052–1064. doi:10.1017/S0950268815002125

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jiang, L., Lee, V., Lim, W., Chen, M., Chen, Y., Tan, L., ... Cook, A. (2015). Performance of case definitions for influenza surveillance. *Eurosurveillance*, *20*(22), 21145.

Jurafsky, D., & Martin, J. (2019). *Speech & language processing*. Pearson Education India.

Kalimeri, K., Delfino, M., Cattuto, C., Perrotta, D., Colizza, V., Guerrisi, C., ... Obi, C., et al. (2019). Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms. *PLoS computational biology*, *15*(4), e1006173. doi:10.1371/journal.pcbi.1006173

Kalousis, A., & Hilario, M. (2000). Supervised knowledge discovery from incomplete data. *WIT Transactions on Information and Communication Technologies*, *25*. doi:10.2495/DATA000261

Katz, R., May, L., Baker, J., & Test, E. (2011). Redefining syndromic surveillance. *Journal of epidemiology and global health*, *1*(1), 21–31. doi:10.1016/j.jegh.2011.06.003

Köpke, K., Prahm, K., Buda, S., & Haas, W. (2016). Evaluation einer ICD-10-basierten elektronischen Surveillance akuter respiratorischer Erkrankungen (SEED ARE) in Deutschland. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, *59*(11), 1484–1491. doi:10.1007/s00103-016-2454-0

Krey, J. (2016). Klinische Ersteinschätzung in der Notaufnahme. *Medizinische Klinik-Intensivmedizin und Notfallmedizin*, *111*(2), 124–133. doi:10.1007/s00063-015-0069-00

Kuhn, M. (2020). *Caret: Classification and regression training*. R package version 6.0-85. Retrieved March 13, 2020, from https://CRAN.R-project.org/package=caret

Larkey, L. S., & Croft, W. B. (1995). *Automatic assignment of icd9 codes to discharge summaries*. Technical report, University of Massachusetts at Amherst, Amherst, MA.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Third edition.). Hoboken, NJ : 1904.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien*. R package version 1.7-3. Retrieved April 5, 2020, from https://CRAN.R-project.org/package=e1071

Mitchell, T. M. (1997). Machine learning. *McGraw Hill*, *45*(37), 870–877.

Možina, M., Demšar, J., Kattan, M., & Zupan, B. (2004). Nomograms for visualization of naive bayesian classifier. In *European conference on principles of data mining and knowledge discovery* (pp. 337–348). Springer.

Olszewski, R. T. (2003). Bayesian classification of triage diagnoses for the early detection of epidemics. In *Flairs conference* (pp. 412–416).

Pakhomov, S. V., Buntrock, J. D., & Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, *13*(5), 516–525. doi:10.1197/jamia.M2077

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2014). Diagnosis code assignment: Models and evaluation metrics. *Journal of the American Medical Informatics Association*, *21*(2), 231–237. doi:10.1136/amiajnl-2013-002159

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved April 5, 2020, from https://www.R-project.org/

Rish, I. (2001). An empirical study of the naive bayes classifier. In *Ijcai 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, *22*, pp. 41–46).

Robert Koch-Institut. (2020a). Arbeitsgemeinschaft Influenza. Retrieved March 11, 2020, from https://influenza.rki.de

Robert Koch-Institut. (2020b). Grippeweb. Retrieved March 11, 2020, from https://grippeweb.rki.de

Scheurwegs, E., Cule, B., Luyckx, K., Luyten, L., & Daelemans, W. (2017). Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, *74*, 92–103. doi:10.1016/j.jbi.2017.09.004

Sentas, P., & Angelis, L. (2006). Categorical missing data imputation for software cost estimation by multinomial logistic regression. *Journal of Systems and Software*, *79*(3), 404–414. doi:10.1016/j.jss.2005.02.026

Sniegoski, C. A. (2004). Automated syndromic classification of chief complaint records. *Johns Hopkins APL Technical Digest*, *25*(1), 68–75.

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, *17*(6), 646–651. doi:10.1136/jamia.2009.001024

Stoto, M. A., Fricker, R. D., Jain, A., Diamond, A., Davies-Cole, J. O., Glymph, C., ... Dehan, K., et al. (2006). Evaluating statistical methods for syndromic surveillance. In *Statistical methods in counterterrorism* (pp. 141–172). doi:10.1007/0-387-35209-0_9

Subotin, M., & Davis, A. (2014). A system for predicting icd-10-pcs codes from electronic health records. In *Proceedings of bionlp 2014* (pp. 59–67).

Torgo, L. (2010). *Data mining with r, learning with case studies*. Chapman and Hall/CRC.

Triple S Project. (2011). Assessment of syndromic surveillance in europe. *The Lancet*, *378*(9806), 1833–1834. doi:10.1016/S0140-6736(11)60834-9

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

WHO Regional Office for Europe. (2020). Seasonal influenza. Retrieved April 1, 2020, from http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/seasonal-influenza

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:10.18637/jss.v059.i10

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved April 5, 2020, from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

World Health Organization. (2018). Influenza (seasonal). Retrieved March 5, 2020, from https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)

World Health Organization. (2020). Classification of diseases (icd). Retrieved April 4, 2020, from https://www.who.int/classifications/icd/en/

Zhang, Y. (2008). A hierarchical approach to encoding medical concepts for clinical notes. In *Proceedings of the acl-08: Hlt student research workshop* (pp. 67–72).

Zhang, Y., Kambhampati, C., Davis, D., Goode, K., & Cleland, J. (2012). A comparative study of missing value imputation with multiclass classification for clinical heart failure data. In *2012 9th international conference on fuzzy systems and knowledge discovery* (pp. 2840–2844). IEEE. doi:10.1109/FSKD.2012.6233805

Zheng, W., Aitken, R., Muscatello, D. J., & Churches, T. (2007). Potential for early warning of viral influenza activity in the community by monitoring clinical diagnoses of influenza in hospital emergency departments. *BMC Public Health*, *7*(1), 250. doi:10.1186/1471-2458-7-250

A Supporting tables and figures

Variable	Description	% missing in whole dataset	% missing in second half
Age*		18.3	1.3
Gender*		< 0.01	< 0.01
PLZ	first three digits of postal code**	0	
Date*		0	0
Hour*		0	0
Department*		18.3	37.4
Disposition	allocation after ED	> 99.9	
Isolation	reason for isolation (if applicable)	46.7	
Referral*	referral to hospital	< 0.01	0.5
Transport	mean of transport to hospital	67.0	
Triage	triage value of severity	44.4	
Vaccination tetanus		46.3	
Complaint (group)*	MTS category	< 0.01	2.3
Complaint (value)*	MTS indicator	< 0.01	3.5
Diagnosis A	excluded diagnosis	> 99.9	
Diagnosis G*	verified diagnosis	58.8	16.1
Diagnosis V	suspected diagnosis	93.3	
Diagnosis Z	diagnosis for condition after sth.	> 99.9	
Pain (healthprofessional)	pain rating by healthprofessional	87.4	
Pain (patient)	pain rating by patient	> 99.9	
Bloodpressure systolic*		55.1	52.1
Heartrate*		55.8	50.6
Oxygensaturation*		57.7	52.3
Respiratory rate*		48.8	10.7
Temperature*		52.3	46.2

* Variable selected for analysis.
** Constant for all visits.

Table A.1: Variables originally included in the dataset. All variables available in the dataset are shown and described where needed, together with their percentage of missing values in the whole dataset (n = 384021). The variables that were selected for the analysis are marked with a star (*).

Variable	Whole set		2nd half w/ missings		2nd half w/o missings*	
	mean	*SD*	*mean*	*SD*	*mean*	*SD*
Age	45.51	26.04	42.13	26.74	56.32	22.49
Temperature (median)	36.95	1.08	36.89	0.96	36.80	0.92
Blood pressure (median)	139.78	24.10	139.50	23.95	139.37	23.84
Oxygen saturation (median)	96.78	3.72	97.07	3.15	96.88	3.17
Heart rate (median)	84.32	21.45	85.61	23.00	83.32	19.05
Respiratory rate (median)	13.84	6.56	13.23	6.30	15.14	5.15
Gender (% female)	48.31		48.09		50.97	

* Complete-case dataset, meaning only visits with no missing values in any of the variables are included. The metrics can of course only be computed for those cases that have an observation in the corresponding variable.

Table A.2: Descriptive statistics for the numeric variables, compared for different datasets. Mean and standard deviation are reported for the numeric variables included in this analysis. For the categorical variable gender the percentage of female patients are shown. They are compared for the different datasets: the whole dataset, including all cases (n = 384021), the second half (n = 188490) and the second half including only complete cases (n = 47088), which is the basis for the training of the models.

referral	Label (German)	Label (English)
VAP	Vertragsarzt/Praxis	panel doctor/doctor's office
KVNPIK	KV-Notfallpraxis am Krankenhaus	emergency doctor's office at the hospital
KVNDAK	KV-Notdienst außerhalb des Krankenhauses	emergency service out of the hospital
RD	Rettungsdienst	ambulance
NA	Notarzt	emergency doctor
KLINV	Klinik/Verlegung	clinic/transfer
NPHYS	Zuweisung nicht durch Arzt	referral not by physician
OTH	Andere	others

Table A.3: Abbreviations in the variable referral.

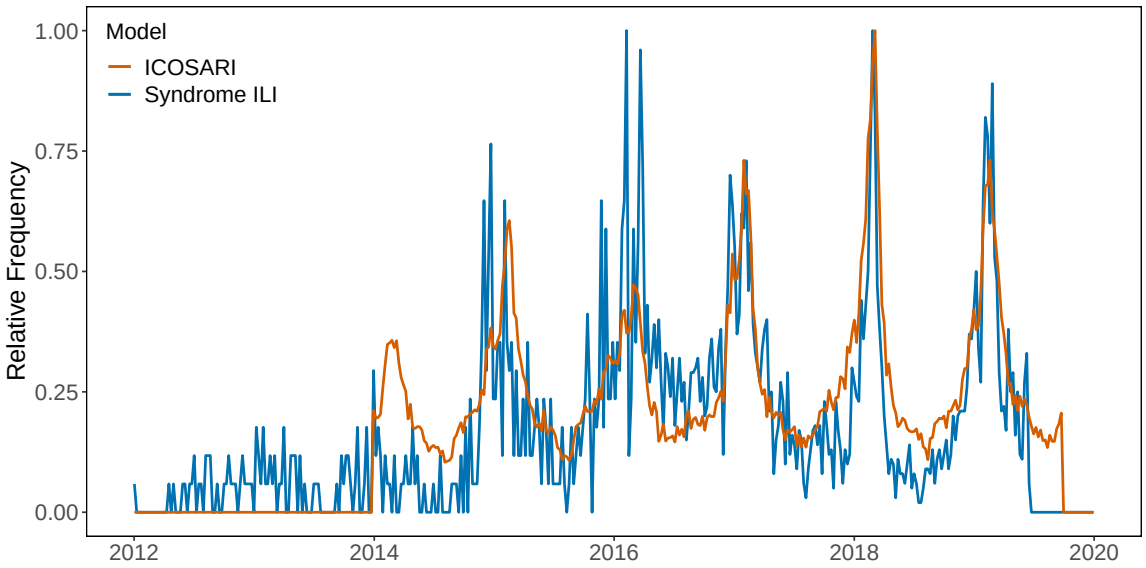


Figure A.1: Weekly aggregated ILI cases compared to ICOSARI data. The cases are scaled to the maximum of each timeseries, resulting in relative frequencies. The whole timeframe is shown, with ICOSARI data being not available before 2014. In the first part, where ICOSARI data is available (2014-01-01 to 2016-03-15) the time series have a correlation of $r = 0.561$ ($p < 0.001$). In the second part, the correlation is $r = 0.786$ ($p < 0.001$).

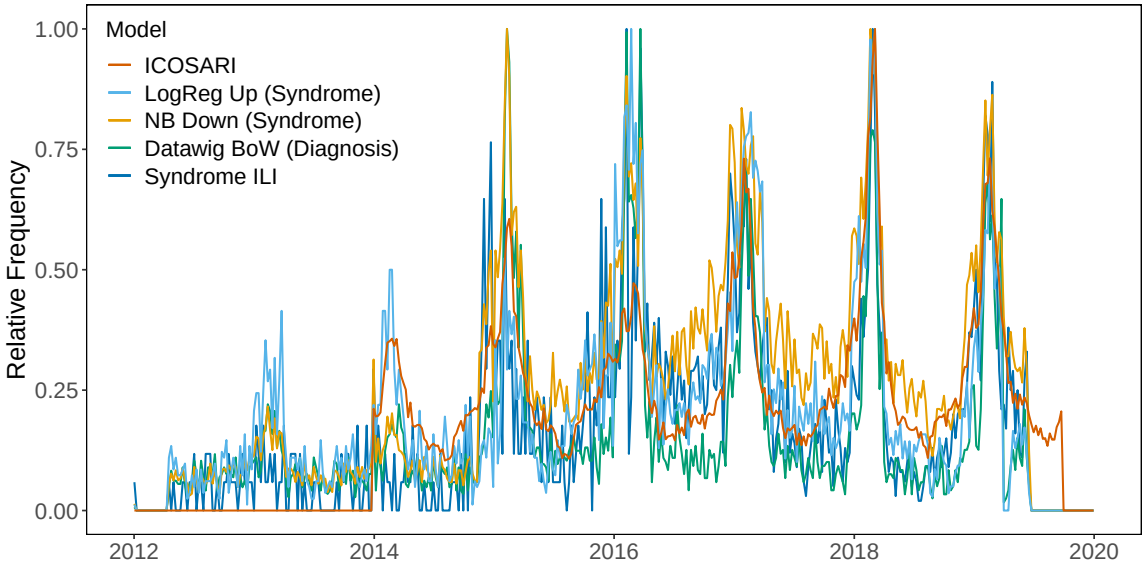


Figure A.2: Weekly aggregated ILI cases by the best models compared to ICOSARI data for the whole time frame. The relative amount scaled to the maximum of each time series is shown, for the whole dataset.

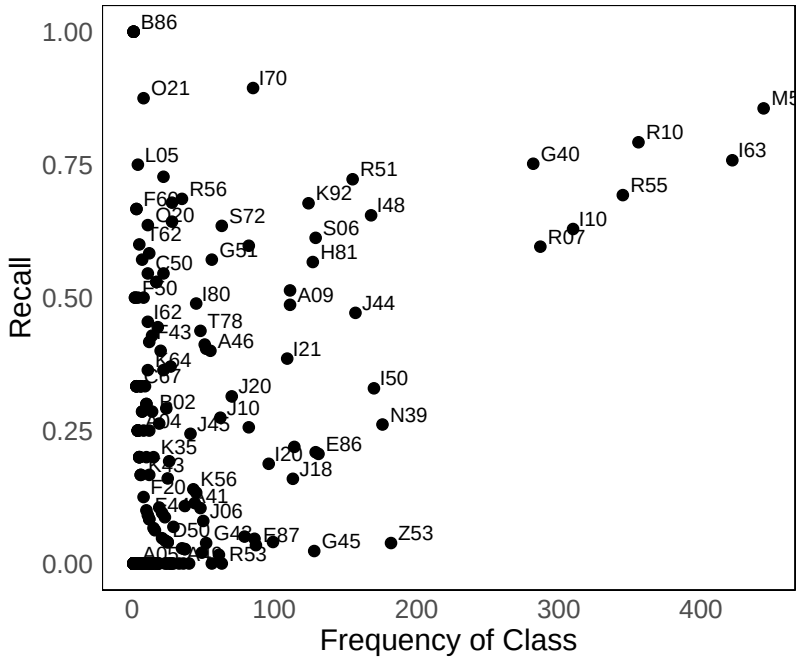


Figure A.3: Recall by class frequency for DataWigBoW. The recall measure for every class (diagnosis (block)) for the imputation model DataWigBoW is plotted against the frequency of this class in the test set.

Package	Version	Package	Version
alluvial	0.1-2	matrixStats	0.55.0
assertthat	0.2.1	mice	3.7.0
broom	0.5.3	naivebayes	0.9.6
car	3.0-7	naniar	0.4.2
caret	6.0-85	plotly	4.9.1
corrplot	0.84	plyr	1.8.5
cowplot	1.0.0	ppcor	1.1
data.table	1.12.8	pROC	1.16.1
dbplyr	1.4.2	pryr	0.1.4
DMwR	0.4.1	RANN	2.6.1
doParallel	1.0.15	readxl	1.3.1
dplyr	0.8.3	rjson	0.2.20
e1071	1.7-3	rlist	0.4.6.1
fst	0.9.0	rmarkdown	2.1
ggalluvial	0.11.1	ROCR	1.0-7
ggplot2	3.2.1	RODBC	1.3-16
ggpubr	0.2.5	rstudioapi	0.10
ggrepel	0.8.1	skimr	2.0.2
gplots	3.0.1.2	stargazer	5.2.2
Hmisc	4.3-0	stringi	1.4.4
ids	1.0.1	tibble	2.1.3
kableExtra	1.1.0	tidyr	1.0.0
klaR	0.6-14	tidyverse	1.3.0
knitr	1.27	timeDate	3043.102
lubridate	1.7.4	VIM	5.1.1

Table A.4: R packages and their versions used in this work.

B Eidesstattliche Erklärung

Ich versichere, dass ich diese Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen habe ich als solche kenntlich gemacht.

Diese Versicherung gilt auch für alle gelieferten Datensätze, Zeichnungen, Skizzen oder grafischen Darstellungen.

Des Weiteren versichere ich, dass ich das Merkblatt zum „Umgang mit Plagiaten“ phoenix.wiwi.uni-bielefeld.de/organisation/pamt/uploads/PlagiatInfo-BlattStudenten.pdf gelesen habe.

(Ort, Datum) (Unterschrift)