MODELING HOSPITAL READMISSIONS IN DALARNA COUNTY, SWEDEN

Authors:
Jiri Prochazka
Sonia Zerin

Supervisors:
Moudud Alam
Hasan Fleyeh
Jonas Gummesson
Mikael Rautio

May 31, 2016

Business Intelligence Program
School for Technology and Business Studies
Dalarna University
Abstract
Unplanned hospital readmissions increase health and medical care costs and indicate lower the quality of the healthcare services. Hence, predicting patients at risk to be readmitted is of interest. Using administrative data of patients being treated in the medical centers and hospitals in the Dalarna County, Sweden, during 2008 – 2016 two risk prediction models of hospital readmission are built. The first model relies on the logistic regression (LR) approach, predicts correctly 2,648 out of 3,392 observed readmission in the test dataset, reaching a c-statistics of 0.69. The second model is built using random forests (RF) algorithm; correctly predicts 2,183 readmission (out of 3,366) and 13,198 non-readmission events (out of 18,982). The discriminating ability of the best performing RF model (c-statistic 0.60) is comparable to that of the logistic model. Although the discriminating ability of both LR and RF risk prediction models is relatively modest, still these models are capable to identify patients running high risk of hospital readmission. These patients can then be targeted with specific interventions, in order to prevent the readmission, improve patients’ quality of life and reduce health and medical care costs.
Contents

1. Introduction ........................................................................................................................................ 1

2. Literature review ................................................................................................................................. 4
   2.1. Length of the follow-up period for hospital readmission ........................................................... 4
   2.2. Methods used .................................................................................................................................. 5
   2.3. Variable used ................................................................................................................................... 6

3. Data and methods ................................................................................................................................. 12
   3.1. Data description ............................................................................................................................. 12
   3.2. Data processing ............................................................................................................................. 13
      3.2.1. Identification of hospitalisations ............................................................................................. 13
      3.2.2. Data reduction related to missing data ..................................................................................... 14
      3.2.3. Reduction of multiple hospital department visits ...................................................................... 15
      3.2.4. Identification of readmissions ................................................................................................. 15
      3.2.5. Data reduction related to lagged information ......................................................................... 15
   3.3. Variable description ....................................................................................................................... 16
   3.4. Proposed methods .......................................................................................................................... 16
      3.4.1. Logistic regression .................................................................................................................... 18
      3.4.2. Random forests ....................................................................................................................... 19

4. Results and Discussion ......................................................................................................................... 23
   4.1. Logistic regression .......................................................................................................................... 23
   4.2. Random forests .............................................................................................................................. 25

5. Conclusion ............................................................................................................................................ 31

References .................................................................................................................................................. 33
1. Introduction

Medical and healthcare related costs are imposing a burden to state budgets, regardless the geographical settings. For instance, in Sweden, the total expenditure on the medical and health care in Sweden in 2013 is estimated to 415 billion SEK, which corresponds to 11% of the Swedish GDP. Moreover, in the period from 2011 to 2013, the share of medical and health expenditure on the total GDP of Sweden increased by 2.5 percentage points (SCB Hälsoräkenskaper, 2013). Similar trend might be observed also in other Nordic countries, in Finland the total expenditure on medical and health care increased from 7.2% of the GDP in year 2011 to 9.4% of the GDP in year 2013. In Denmark the corresponding number of health care expenditure in year 2011 is 8.7% of the GDP, while in 2013 the share increased to 11.3% (WHO Global Health Observatory, 2015).

Among the many cost drivers of medical and health care, hospital care is one of the major contributors. One part of the hospital care is represented by readmissions, i.e. repeated admissions of patients within a certain time frame following a hospitalization. Particularly in the United States, hospital readmission receives interest both among policy makers and researchers and represents a frequently used metrics for comparisons of the quality of healthcare provided. Jencks, Williams, and Coleman (2009) reported that in the United States, in the period 2003-2004 almost one fifth (19.6%) of the 11,855,702 patients eligible to the Medicare were repeatedly admitted to the hospital within 30 days following the initial hospitalization, and that 34.0% of the patients were readmitted within 90 days. In a Swedish study, Edgren, Stäck, Forsberg, Högberg, and Werr (2012) showed that 1% of the most resource-intensive patients are admitted in average four times a year and that admissions of this patient group accounts for 25% of all admissions annually.

Many of the hospital readmissions are unavoidable, e.g. patients are re-hospitalised as a follow-up of the initial admission or for a newly emerged disease or disorder, which is unrelated to the initial hospital stay. However, as shown in several studies, certain part of the hospital readmissions are unplanned one, results of complications, such as deviations from the prescribed medication, inadequate discharge planning or failure to seek medical care once worsening symptoms appear (e.g. Ekman et al. 1998). These readmissions could be avoided, given that the patients would be targeted with relevant interventions during the initial hospitalization and in the follow-up period. The share of avoidable hospital readmission on all readmission varies in the literature; the exact proportion depends on the population of patients studied, length of the follow-up period and methodology of the study and in particular, on the criteria to define the avoidable readmission. Jencks et al. (2009), relying on a dataset of Medicare beneficiaries, assessed that 78% of all readmission in their data could be avoided, given the patients would be targeted with interventions during their hospital stay. On the other hand, van Walraven, Bennett, Jennings, Austin, & Forster (2011) in a systematic review of 34 studies found...
that the percentage of readmission considered as avoidable ranged between 5% and 78.9% with a median value 27%. In another review study, Yam et al. (2015) assess the prevalence of avoidable readmission to vary between 9% and 59%. Regardless the exact share of the avoidable readmission, there is a clear consensus in the literature that unplanned hospitalizations should be reduced, since these are related with unnecessary expenditure and lower quality of healthcare (e.g. Jencks et al., 2009).

Hospital discharge process and care transition are some of the topics receiving attention in the literature. Interventions related to these are believed to reduce the share of avoidable hospital readmission. Possible interventions are for instance nursery visits at patient’s home (Chiu & Newcomer, 2007; Hall et al., 2014; Jackson, Kasper, Williams, & DuBard, 2015) and follow-up telephone calls (Harrison, Hara, Pope, Young, & Rula, 2011). Provision of sufficient discharge information to the patients is another identified intervention (e.g. Coleman, Parry, Chalmers, & Min, 2006; Jack et al., 2009), along with educating household members in order to provide informal post hospital care (e.g. Armour, Norris, Jack, Zhang, & Fisher et al., 2005) and identifying and educating of care givers outside the patients’ household (e.g. Piete et al., 2015; Rodriguez-Gonzalo, Garcia-Marti, Ocan-Colorado, Baquera-De Michelo, & Morel-Fernandes, 2015). These scientific studies bring evidence that aforementioned interventions in certain extent reduce the prevalence of hospital readmissions and improve the overall quality of patient care. For instance, Harrison et al. (2011) reports that patients in a control group who received a follow-up call after hospital discharge were 23.1 % less likely to be readmitted within 30 days, compared to the comparison group, who did not receive a follow-up call. Hall et al. (2014) show that intervention in form of visits by the nurses at patients’ home resulted in 3.85% readmission rate for the treatment group, while in a comparable control group, the share of readmissions was 11.54 %.

Many of the above mentioned post hospital interventions are personal- and resource intensive, if applied to all discharge patients. In order to reduce the overall costs of the interventions and maximize the benefits, readmission risk models were developed. These prediction models, using administrative and patient data, enables to identify patients running high risk of hospital readmission already through the initial hospital stay. The readmission reducing interventions can then be applied to specific high-risk population, i.e. to those patients who would benefit most, rather than to be aimed on patients who are not at high risk.

Readmission risk modelling is rather common in the United States (e.g. Coleman, Min, Chomiak, & Kramer, 2004; Krumholz et al., 2000; Krumholz et al., 2011); in Europe they are used very scarcely (Bottle, Aylin, & Majeed, 2006 in UK; Halfon et al., 2006 in Switzerland). To our best knowledge, there is only one recent study building a predictive model using Swedish data (Edgren et al., 2012). This study aimed at identifying patients running high risk of readmissions in the county of Stockholm.
With the practical impact of the predictive models of hospital readmission on reduction of medical and healthcare costs and the relative scarcity of Swedish studies in mind, the aim of this study is to develop a predictive model that enables to estimate hospital readmission risk of patients treated in the hospitals in Dalarna County, Sweden, with a high precision given the patients characteristics, and other treatment oriented factors.

To achieve the aim, we reviewed the existing literature focused on predictive modelling of hospital readmission risk, in order to identify suitable methods to build predictive model(s). Other aims of the literature review were to identify optimal length of follow-up period for readmission definition and factors associated with patients running high risk of readmission. Based on the administrative data provided by Landstinget Dalarna, we build suitable predictive models, as per the state of art in the subject area. The discrimination ability of the proposed models is also assessed.

The rest of this thesis is structured as follows: In Chapter 2, the most recent studies on risk prediction modelling are reviewed, with main emphasize on the length of the follow-up period to measure hospital readmission, methods and explanatory variables being used in risk prediction modelling. Chapter 3 describes the data, along with the process of data acquisition and pre-processing. The proposed prediction model classes, the logistic regression and the random forest are presented in Chapter 3 as well. Chapter 4 reports the results, while Chapter 5 concludes the thesis.
2. Literature review

Studies on the risk prediction of hospital readmission are abundant. The main focus of this literature review is to identify a common length of a follow-up period for the definition of hospital readmission, the analytical methods being found successful, and predictive/explanatory variables being commonly used in hospital readmission modelling.

Table 1 and Table 2 summarize different studies on hospital readmission risk, considered in this literature review. When searching the most relevant papers, two top cited review studies were selected (Kansagara et al., 2011; van Walraven et al., 2011). Studies reviewed in both of them made it to the final selection, but just those explicitly describing data and methods used in the study, where studies using only clinical data were excluded. Also most recently published studies about prediction of hospital readmission, from authors discussed in both reviews, were considered, under the same conditions about the data and methods. Most of the studies, 11 out of 15, used US medical data. One study (Bottle et al., 2006) used data from the UK, one study (van Walraven et al., 2010) used data from Canada, while two studies used data from Australia (Howell, Coory, Martin, & Duckett, 2009) and New Zealand (Futoma, Morris, & Lucas, 2015) respectively. There is rather large variation in the size of the sample used in the reviewed studies, ranging from a study based on a dataset of 499 patients (Natale, Wang, & Taylor, 2013) to three studies, Bottle et al. (2006), van Walraven et al. (2010) and Futoma et al. (2015) with more than 1 million patients. The studies also differ with respect to the settings and population. In some studies, hospital readmission was predicted in general population settings (Halfon et al., 2006; Hasan et al., 2010; Howell et al., 2009) while other studies restricted the analysis to a specific patient population, such as patients over 65 years (Hammill et al., 2011; Krumholz et al., 2000; Silverstein, Qin, Mercer, Fong, & Haydar, 2008) or specific diagnosis, i.e. congestive heart failure (Amarasingham et al., 2010; Hammill et al., 2011; Natale et al., 2013) or surgical patients (van Walraven et al., 2010).

2.1. Length of the follow-up period for hospital readmission

In the literature, the hospital readmission is commonly defined as an unplanned admission to a hospital within a certain time frame (generally 30 days), following an original hospital stay. A readmission can occur both in the same hospital as the original stay, or in a different hospital or other medical facility. Besides 30 days, a few other time frames are also adopted. Some studies used a relatively long time span of the follow up period such as 6 months (Krumholz et al., 2000) and 12 months (Bottle et al., 2006; Howell et al., 2009). Fialho et al. (2012) classify patients as being readmitted if the patient is admitted back to the hospital between 24 hours and 72 hours after the initial discharge.
2.2. Methods used

With respect to the methods being used for the readmission risk modelling, the reviewed studies can be grouped into two categories - studies relying on (classical) statistical methods (Table 1) and studies using various data mining approaches (Table 2). Compared to the statistical methods, data mining approaches are used relatively scarcely.

Among the studies using statistical method, the dominating method being used is the logistic regression which is used in 8 out of 10 reviewed studies which used statistical methods. Some studies (e.g. Amarasingham et al., 2010; Bottle et al., 2006; Silverstein et al., 2008) used simple logistic regression, in order to assess the impact of various risk factors on hospital readmission as well to make prediction of readmission risks. The simple logistic regression is frequently used in studies based on homogenous population, such as patients of a single hospital or patients of several hospitals, sharing the same characteristics.

Lindenauer et al. (2011) and Hammill et al. (2011) used hierarchical logistic regression with a random hospital-specific intercept, in order to account for within-hospital correlation and to model the possible underlying differences in the quality between hospitals that might lead to systematic differences in hospital readmissions. According to these authors this modelling approach enables to estimate hospital-specific comparison rate, which may be used to assess hospital performances. Van Walraven et al. (2010) use a technique of generalized estimating equation in order to account for clustering of patients within hospitals, i.e. the fact that some of the variables measured on the hospital-level might be similar for more patients. Hasan et al. (2010) deploy generalized estimating equation in order to account for the fact that the discharged patients were treated by different physicians and hospital sites were entered as fixed effects, in order to minimize confounding. Besides the logistic model (and its extensions), Cox’s proportional hazard model is also used (Krumholz et al., 2000). Among the reviewed studies, the data mining approaches varies from decision trees (Natale et al., 2013), random forest (Zolfanger, Meadem, Teredesai, Roy, & Chin, 2013), fuzzy modelling (Fialho et al., 2012) to support vector machines (Yu, Farooq, Esbroeck, Fung, & Anand, 2015). Futoma et al. (2015) compared several predictive models, some of which were not previously applied in predictive modelling of hospital readmission. In this comparison, among the data mining approaches, model based on random forests were found to have the highest prediction ability.

To assess the predictive power of the prediction models, mainly, the so called ROC statistics is used. In Bottle et al.’s (2006) words: “The receiver operating characteristic (ROC) or c-statistics is widely used to summarize a model’s ability to correctly discriminate between outcomes such as whether the patient will be readmitted. A value of c = 0.50 suggests that the model is no better than chance in predicting readmission, while a value of 1 suggests a perfect discrimination. In general, values less
than 0.70 are considered to show poor discrimination, whereas values above 0.80 suggest very good discrimination”.

Among the reviewed studies reporting the c-statistic, values stretch from \( c = 0.60 \) to \( c = 0.72 \). In five studies, the value of the c-statistics were over \( c = 0.7 \), which indicates modest discriminating power of the models. Two of these studies (having \( c \geq 0.70 \)) used statistical methods. Amarasingham et al. (2010) studied a 30-day readmissions risk of heart failure patients in a major urban hospital and after controlling for indicators of lower socio-economic status of the patients achieved a predictive power of \( c = 0.72 \). Bottle et al. (2006) is the other study relying on statistical models, where the authors achieved a c-statistic also equal to \( c = 0.72 \) in a study using population of patients to emergency departments of UK hospitals, and predicting high-risk of emergency readmission in the 12 months following the index admission. The other three studies with c-statistics over the value 0.70, used data mining approach.

2.3. Variable used

Generally, the goal of the risk prediction models is to identify patients, running high risk of hospital readmission, once accounting for a variety of factors, expected to be linked to the readmission risk. Kansagara et al. (2011) suggest categorizing the risk factors used in the prediction models into four broad categories:

i) factors describing medical comorbidity,

ii) overall health conditions,

iii) prior use of medical services, and

iv) sociodemographic and socioeconomic factors.

The first group of factors encompasses factors related to the so-called comorbidity. Comorbidity means in medicine a presence of additional diseases or disorders related to the primary diagnosis. Generally, the idea behind including comorbidity in the prediction models is that a presence of additional diseases is linked to a higher risk of hospital readmission. In the risk prediction studies, it may be found a broad variety of additional diseases and disorders, which the authors control for. The types of comorbidities which are considered depend commonly on the aim and scope of the particular study. For example, Tsai, Joynt, Oray, Gawande, and Ashish (2013) in a recent study analyzed readmission rates related to major surgeries, such as coronary-artery bypass and hip replacement of 479,471 Medicare patients discharged from 3,004 hospitals in the USA. In this study, the comorbidities being controlled for are coexisting diseases, on which the patients suffer, such as
congestive heart failure, chronic obstructive pulmonary diseases, hypertension, diabetes mellitus, obesity or depression. Another way how to account for the effects of comorbidity is with a help of so called comorbidity index. Hasan et al. (2010) uses a Charleson comorbidity index, when for each severe disease the patient suffers from, such as cancer, AIDS or severe liver disease, a certain number of points is added to a total score. The authors report the Charleson comorbidity index to be an important predictor of future readmission. In van Walraven et al. (2010) study, Charleson comorbidity index is adopted as well. Similarly to Hasan et al. (2010), van Walraven et al. (2010) report that the Charleson comorbidity index is significantly associated with a risk of hospital readmission.

The second group of factors is connected with overall health conditions. Illness severity is one of frequently used factors to describe health conditions. Illness severity can be measured with severity index or laboratory findings. The actual measures employed vary with the aim and scope of the analysis under scrutiny. In a study of risk for 30-day readmission of heart failure patients, Amarasingham et al. (2010) use severity of illness score developed by Tabak, Johannes, and Silber (2007), which is based on the age of the patient and 17 vital sign and laboratory variables, measured within the first 24 hours of patients’ hospital admission. The authors found that the inclusion of this severity index provides a strong and statistically significant predictor of readmission in the regression analysis. Another way how to control for illness severity is to deploy laboratory findings. In this sense, Hammill et al. (2011) use laboratory results such as serums of creatine and sodium, hemoglobin and a value of systolic blood pressure and reports that these laboratory results have statistically significant impacts upon readmission rate of patients with a heart failure diagnosis. Type of diagnosis on which the patient suffers is another factor frequently included in readmission risk modelling (e.g. Krumholz et al., 2011).

Previous use of medical services, measured in the period prior to the initial hospitalization is another type of factor frequently accounted for in readmission risk modelling. Bottle et al. (2006) in a study analyzing 2,895,234 patients admitted as emergencies into the hospitals in the UK controlled for a number of visits to hospitals’ emergency departments in the 365 days long period prior to the index admission. The authors found that there is a strong relation between visiting emergency hospital departments and subsequent hospital readmission, nearly one half of the patients, who had three and more emergency visits in the previous 365 days, went on subsequent readmission. Hasan et al. (2010), in a study of readmission rates among general medicine patients hospitalized in six large academic hospitals in the USA, assess the previous use of medical services as a number of hospitalization in last year prior to the hospitalization. Also this study reports that previous number of hospitalizations to be an important prediction factor of subsequent hospitalization. Other ways how to measure patient’s use of medical services prior to the hospitalization are clinic visits (e.g. Amarasingham et al., 2010; Billings & Mijanovich, 2007) and hospital length of stay (Hasan et al., 2010; van Walraven et al.,
Table 1: Studies on hospital readmission using statistical methods

<table>
<thead>
<tr>
<th>Authors</th>
<th>Settings</th>
<th>Country</th>
<th>Sample size</th>
<th>Length of follow-up period</th>
<th>Model</th>
<th>c-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krumholz et al. (2000)</td>
<td>Medicare patients aged 65 + with heart failure, treated in multiple medical centres in one US state, 1994-1995</td>
<td>USA</td>
<td>2,176</td>
<td>6 months</td>
<td>Cox proportional hazard model</td>
<td>Not reported</td>
</tr>
<tr>
<td>Krumholz et al. (2011)</td>
<td>Medicare patients with acute myocardial conditions in the year 2006 year</td>
<td>USA</td>
<td>200,750</td>
<td>30 days</td>
<td>Hierarchical Logistic regression</td>
<td>0.63</td>
</tr>
<tr>
<td>Bottle et al. (2006)</td>
<td>All patients in period 2000-2001</td>
<td>UK</td>
<td>2,895,234</td>
<td>12 months</td>
<td>Logistic regression</td>
<td>0.72</td>
</tr>
<tr>
<td>Silverstein et al. (2008)</td>
<td>Patients 65 years and older, treated in several centres in a single city from 2002 to 2004.</td>
<td>USA</td>
<td>29,292</td>
<td>30 days</td>
<td>Logistic regression</td>
<td>0.65</td>
</tr>
<tr>
<td>Howell et al. (2009)</td>
<td>Public hospital patients, with chronical medical condition, discharged 2005-2006</td>
<td>Australia</td>
<td>17,699</td>
<td>12 months</td>
<td>Logistic regression</td>
<td>0.65</td>
</tr>
<tr>
<td>Amarasingham et al. (2010)</td>
<td>Hearth failure patients treated in urban hospital 2007-2008</td>
<td>USA</td>
<td>1,372</td>
<td>30 days</td>
<td>Logistic regression</td>
<td>0.72</td>
</tr>
<tr>
<td>Hasan et al. (2010)</td>
<td>Patients of six academic hospitals</td>
<td>USA</td>
<td>10,946</td>
<td>30 days</td>
<td>Logistic regression</td>
<td>0.61</td>
</tr>
<tr>
<td>van Walraven et al. (2010)</td>
<td>Medical and surgical patients treated in 11 Canadian hospitals</td>
<td>Canada</td>
<td>1,004,812</td>
<td>30 days</td>
<td>Logistic regression</td>
<td>0.68</td>
</tr>
<tr>
<td>Hammill et al. (2011)</td>
<td>Patients aged 65 + with heart failure discharged 2004-2006</td>
<td>USA</td>
<td>24,163</td>
<td>30 days</td>
<td>Generalized linear regression</td>
<td>0.60</td>
</tr>
<tr>
<td>Lindenauer et al. (2011)</td>
<td>Medicare patients 65+ with pneumonia, general US population</td>
<td>USA</td>
<td>226,545</td>
<td>30 days</td>
<td>Logistic regression</td>
<td>0.63</td>
</tr>
<tr>
<td>Authors</td>
<td>Settings</td>
<td>Country</td>
<td>Sample size</td>
<td>Length of follow-up period</td>
<td>Model</td>
<td>c-statistic</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>---------</td>
<td>-------------</td>
<td>----------------------------</td>
<td>-------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Futoma et al. (2015)</td>
<td>All patients admitted to New Zealand hospitals between 2006 and 2012</td>
<td>NZ</td>
<td>3,295,775</td>
<td>30 days</td>
<td>Random forests</td>
<td>0.68</td>
</tr>
<tr>
<td>Natale et al. (2013)</td>
<td>Hearth failure patients treated between August 2011-May 2012</td>
<td>USA</td>
<td>499</td>
<td>30 days</td>
<td>Decision tree</td>
<td>Not reported</td>
</tr>
<tr>
<td>Zolfaghar et al. (2013)</td>
<td>Multicare patients with heart failure patients, treated in the period 2009-2013</td>
<td>USA</td>
<td>1,372</td>
<td>30 days</td>
<td>Random forests</td>
<td>0.72</td>
</tr>
<tr>
<td>Fialho et al. (2012)</td>
<td>Patients aged 15+, ICU inpatients</td>
<td>USA</td>
<td>1,028</td>
<td>72 hours</td>
<td>Fuzzy modelling</td>
<td>0.72</td>
</tr>
<tr>
<td>Yu et al. (2015)</td>
<td>Medicare/Medicaid patients of three hospitals, aged 65+</td>
<td>USA</td>
<td>26,520</td>
<td>30 days</td>
<td>Support vector machine</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Sociodemographic variables such as patient’s age are considered as predictors in the majority of studies. Generally, age is reported to have a weak positive linear relationship to hospital readmission, in other words, with increasing age of the patients the readmission rate is higher. Gender is another frequently used characteristic in risk prediction models. Like the age, also gender is frequently found to be rather weak predictor of the readmission (Billings & Mijanovich, 2007; Holman, Preen, Baynham, Finn, & Semmens, 2005). In some of the reviewed studies (e.g. Hasan et al., 2010; Krumholz et al., 2000; van Walraven et al., 2010), authors included age and gender as explanatory variables in the initial models being tested, however, were excluded from the final prediction models due to their rather weak discriminating ability.

Lastly, the risk prediction studies control also for various individual characteristics, describing the socio-economic factors such as patient’s income (Bottle et al., 2006; Holman et al., 2005; Howell et al., 2009), employment status (Amarasingham et al., 2010; Billings & Mijanovich, 2007), education level (Hasan et al., 2010; Morrissey et al., 2003), availability of care givers, access to medical care (e.g. Hasan et al., 2010; Howell et al., 2009) and area-level deprivation (Bottle et al., 2006). In studies analyzing readmission rates of elderly patients, a factor being commonly controlled for is the discharge location, i.e. whether the patient is discharged to home or nursing home (see e.g. Silverstein et al., 2008).

In order to select variables to be included in the risk-prediction models, pre-modelling commonly starts with a set of variables, informed by literature review, expert knowledge and data availability. Several approaches are used in order to select predictive variables to be included in the final models. Stepwise (forward and backward) regression is one of the techniques, frequently used in the reviewed studies listed in Table 1. This technique implies that in the initial model(s), available variables are tested, while variables below certain threshold are then included / excluded from the final model. Exact technique and inclusion/exclusion threshold varies between studies. Howell et al. (2009) used backward elimination and considered variables as candidates for the model if they were univariately significant at $P < 0.25$ and variables were then retained if they remained significant at $P < 0.10$. Krumholz et al. (2000) used an entry level $P<0.20$ and exit level $P>0.10$. Hammill et al. (2011) used a specific selection method, in order to evaluate the incremental value of including clinical data for readmission prediction. Using 500 bootstrapped samples of the data, backward selection is used in order to select the candidates among a large number of available clinical variables. Variables with $P < 0.05$ are kept in the bootstrapped models, while to be included in the final model, the variables need to be retained in at least 60% of the bootstrapped 500 models, i.e. in 300 models. Silverstein et al. (2008) combines forward and backward technique. Forward addition is used for each class of covariates to
evaluate additional contribution to the model. Models using all the variables were further analyzed using backward elimination to identify a parsimonious set of covariates significant at \( P<0.05 \). Van Walraven et al. (2010) is using backward stepping with a \( P < 0.01 \) to include only significant variables in the final multivariable logistic regression model. The authors argue that the rather strict criterion of \( P < 0.01 \) is used in order to ensure model stability.

To summarize, the literature review reveals that the most frequently used length of the follow up period for definition of hospital readmission is 30 days, which was used in eleven of fifteen reviewed studies. Furthermore, two types of methods are utilized in studies of readmission risks modelling – statistical methods and data mining approaches. Logistic regression (LR) is by far the most frequently appearing method utilized for model building in studies based on statistical methods. The scope of studies utilizing various data mining techniques is substantially smaller once compared to the studies using logistic regression. The risk prediction models based on data mining approaches emerged relatively recently. It is interesting to note that there seems not to be any difference in the discriminating ability, measured with c-statistics, of the models based on statistical methods and data mining approaches. Varieties of factors, included in the prediction models, range from socioeconomic characteristic of the patients to detailed clinical results. In order to evaluate the importance of explanatory factors / variables to be included in the final prediction models, technique of forward and backward discrimination is commonly utilized.
3. Data and methods

3.1. Data description

This study relies on administrative data of patients being provided medical care in hospitals and primary care centres located in Dalarna County, Sweden. The data cover the period between January 2008 and January 2016, and were provided by Landstinget Dalarna.

Figure 1 illustrates the structure of the dataset provided by Landstinget Dalarna. The main (master) table in the database is labelled as “Care occasions”. It consists of 32 million observations, where one observation corresponds to one unique medical visit of a patient. In other words, a person being provided medical care in one of the hospitals and other medical centres belonging to the Landstinget Dalarna generates one row in this table. For a patient being transferred during one hospital visit to several hospital departments, several rows are generated in this table, each row then corresponds to medical care provided by one particular hospital department visited.

![Figure 1: Structure of the data base provided by Landstiget Dalarna](image)

The master table might be related through a primary key to three additional (child) tables labelled as “Diagnosis”, “Medication”, and “Treatment”. The “Diagnos” table contains information related to patients’ diagnosis, the “Medication” table contains information about all pharmaceuticals and other medicine prescribed to the patient, while the “Treatment” table contains information about all the medical actions prescribed to the patients. Thus, for each patient record in the “Care occasions” table, the main diagnosis may be identified for the particular medical visit in table “Diagnosis”, what medicine or other pharmaceutics were prescribed in table “Medication”, and what medical actions were taken, in table “Treatment”. In order to ensure privacy, the primary key was for the purpose of this work replaced by new, encrypted one.
The data in the tables “Care occasions”, “Diagnosis”, “Medication” and “Treatment” were collected through three different systems – “Elvis”, “J3” and “TakeCare (TC”) Out of the 32 million rows in the “Care occasions” table, 4.5 million rows were collected through the currently used TC system, roughly 23 million rows were collected through the J3 system, and through Elvis, around 4 million rows were collected.

These three different systems differ in the amount and scope of data being collected. For the purpose of this work, only admissions captured through the TC system could be utilized. The main reason for not being able to use data gathered through J3 is the fact that this system was utilized in order to collect data from medical visits to primary care centres, i.e. medical care provided without hospitalization. The main reason for not being able to use the data gathered by the Elvis system, is non-compatibility with the current system, TC, in particular, the definition of hospital admission is substantially different in Elvis in comparison to the current system TC. However, the information from the two other systems – J3 and Elvis – were used to gather some supplemental information, e.g. chronical diseases, as explained further in the text.

3.2. Data processing

In order to build the predictive models of hospital readmission risk, the dataset provided by the Landstinget Dalarna were pre-processed in several steps, using SQL Server 2016®. The goal of this data pre-processing procedure is to clean the data, to select part of the data relevant for the predictive modelling, and to define the dependent variable in the analysis – hospital readmission, along with the explanatory variables. This procedure was carried out through the following five steps as shown in Figure 2.

3.2.1. Identification of hospitalisations

The first step in the data pre-processing procedure is to select observations being collected through the TC system, which corresponds to approximately 4.5 million rows in the “Care occasions” table. Furthermore, the “Care occasions” table contains records related to hospital admissions, as well records of patients’ visits in primary care centres and of medical care provided in hospitals without hospitalization. Given the aim of the study, which is to predict hospital readmission, rows representing patients’ visits to primary care centres, as well as medical care without hospitalization were filtered out, i.e. only inpatients were kept in the dataset. The data reduction in this step resulted in 157,804 rows.
3.2.2. Data reduction related to missing data

In the second step, rows with missing or incorrect information related to the date of hospital admission and hospital discharge were removed from the “Care occasions” table. Due to missing hospital discharge date, 617 rows were removed. Another 234 rows were removed for having the discharge date dated earlier than admission date. In addition to the missing and incorrect dates of hospital admission/discharge, 597 rows with missing information on whether the patients’ admission was acute or not were removed from the dataset. The information on the acute/not acute admission is crucial, since this is one of the factors defining hospital readmission, as explained later. Furthermore, admission records of patients not resided in the Dalarna county, which represent 10,508 rows were deleted because the next hospital readmission of patients from other counties would most likely to occur in their home county, which means that any readmission incidence of those patients would not
be identifiable from the dataset provided by Landstinget Dalarna. After the second pre-processing step the size of dataset reduced to 141,445 rows.

3.2.3. Reduction of multiple hospital department visits
In the third step, the cases where a patient during a hospital visit is transferred through several hospital departments are handled. As mentioned above, if a patient was provided medical care by three hospital departments, those records would appear in three rows in the “Care occasions” table, each row corresponding to a medical care provided by one particular hospital department. However, for the purpose of the analysis, such multiple hospital department visits, yielding several rows, need to be reduced to a single row only. After this step the size of the dataset further reduced to 128,920 rows.

3.2.4. Identification of readmissions
In the fourth step the hospital readmissions, i.e. the dependent variable for the analysis, is identified. For each row in the “Care occasions” table, it must be identified whether the respective row in the “Care occasions” table represents a hospital readmission or not. The definition of the readmission used in this study is straightforward. A readmission is when following two criteria are fulfilled (1) the patient’s admission is marked as acute and (2) the patient is admitted within 30 days after the discharge related to the initial hospital admission. Thus, in this step, the rows in the “Care occasions” table were sorted in ascending order with respect to the patients’ identification key and hospital admission date. Then, two consecutive rows are compared and given that these rows contain identical patient’s identification key, the time gap between the discharge date in the first row and the admission date in the second row is less than 30 days and the admission in the second row is marked as acute, then the incident presented in the latest row is identified as hospital readmission.

3.2.5. Data reduction related to lagged information
In the fifth step, cases when the patient passed away during a hospitalization are removed. Furthermore, rows, representing patients’ last admissions with discharge date not older than 30 days from the date of data acquisition for this work, were removed, along with rows when patient dies within 30 days after the last admission. In these two cases, it is not possible to see whether the next admission occurs within 30 days or not, and thus, the value of the dependent variable cannot be assigned to these rows. The very final, but quite considerable reduction of the dataset is due to the fact that a variable, counting number of previous readmissions within the last year, is used. For this reason the data coming from the first year of use of TC cannot be used and approximately 1/3 of the data were removed. The final dataset comprises of 74,492 rows. Besides the pre-processing of the data stored in the “Care occasions” table, tables “Läkemedel” and “Diagnosis” were used to retrieve explanatory variables for the analysis. These variables are described in following chapter.
3.3. Variable description

In order to be able to build the prediction model for hospital readmission, several variables of interest were retrieved from the tables “Diagnosis” and “Medication”, other variables were calculated using the information from the table “Care occasions”. The choice of the variables to be evaluated in the modelling part of this study is based on the literature review, presented in Chapter 2 and influenced by data availability.

Dependent variable in the analysis is a binary variable, NextReadm, which is equal to one given a patient is acutely admitted to the hospital within 30 days following a discharge, and zero otherwise.

Explanatory variables are grouped into several groups – sociodemographic variables (Age, Gender, and Municipality), overall health (Diagn, Severity, Antibio), prior use of medical services (Readmission, PrevReadm and LengthOfStay), and comorbidity (ChronicalDis and Antibio). It is also considered at what hospital department is the patient hospitalized (Department) and the financial demandingness of the medical treatment (TreatDem), which is grouped into the last category, miscellaneous. Several other variables would be of interest for this study, such as the socio-economic status of the patients, however, these were not available to the authors.

Table 3 provides an overview over the variables being considered in the analysis, including the type of the variable, mean and standard deviation (for numerical variables), as well as variable description and source. In the final dataset, there are 14.9 % of readmissions. Thus, it can be observed that the final dataset is unbalanced, with respect to the readmissions.

3.4. Proposed methods

The literature review (see Chapter 2) revealed that the majority of predictive models of hospital readmission rates used logistic regression. In addition to the logistic regression, the literature review revealed that in recent decade, also several methods within the data mining approach (e.g. support vector machines, fuzzy logic, random forests) were applied in the readmission risk modelling, achieving comparable prediction performance as models based on logistic regression. Therefore, this study will build two risk prediction models, first model relying on logistic regression (LR), second model based on random forests (RF). There are two main reasons why the random forests method is selected for this research, amongst the other data mining tools. Firstly, the literature review showed that RF is an efficient classifier to predict the hospital readmission. Secondly, unlike some other widely used data mining methods (e.g. neural network or support vector machine), random forests is not a “black box”, which means that apart from the classification itself, this technique provides also information about the importance of the particular variables for the outcome.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Mean</th>
<th>Std.D</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NextReadm</td>
<td>binary</td>
<td>0.15</td>
<td></td>
<td>Is the subsequent visit</td>
<td>Care occasions/</td>
</tr>
<tr>
<td></td>
<td>(1=readmission, 0=otherwise)</td>
<td></td>
<td></td>
<td>readmission or not</td>
<td>Own calculation</td>
</tr>
<tr>
<td><strong>Sociodemographic and socioeconomic variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>numerical</td>
<td>58.60</td>
<td>24.30</td>
<td>Age of the patient</td>
<td>Care occasions</td>
</tr>
<tr>
<td>Gender</td>
<td>factor</td>
<td>0.54</td>
<td></td>
<td>Gender of the patient</td>
<td>Care occasions</td>
</tr>
<tr>
<td></td>
<td>(0=male, 1=female)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Municipality</td>
<td>factor</td>
<td></td>
<td></td>
<td>Municipality where the patient has</td>
<td>Care occasions</td>
</tr>
<tr>
<td></td>
<td>(15 levels)</td>
<td></td>
<td></td>
<td>residence</td>
<td></td>
</tr>
<tr>
<td><strong>Overall health</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td>factor</td>
<td>0.67</td>
<td></td>
<td>Level of severity of the diagnoses</td>
<td>Care occasions</td>
</tr>
<tr>
<td></td>
<td>(5 levels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibio</td>
<td>binary</td>
<td></td>
<td></td>
<td>Use of antibiotics during the admission</td>
<td>Medication</td>
</tr>
<tr>
<td></td>
<td>(1=ATB prescribed, 0=otherwise)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagn</td>
<td>factor</td>
<td></td>
<td></td>
<td>Diagnosis group related to the current admission based on ICD-10 coding</td>
<td>Diagnosis</td>
</tr>
<tr>
<td></td>
<td>(22 levels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prior use of medical services</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PrevReadm</td>
<td>numerical</td>
<td>0.34</td>
<td>1.15</td>
<td>Number of previous readmissions within 365 days</td>
<td>Care occasions/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Own calculation</td>
</tr>
<tr>
<td>LengthOfStay</td>
<td>numerical</td>
<td>4.86</td>
<td>9.05</td>
<td>Length of stay in the hospital in days</td>
<td>Care occasions/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Own calculation</td>
</tr>
<tr>
<td>Readmission</td>
<td>binary</td>
<td>0.14</td>
<td></td>
<td>Is the current visit readmission or not</td>
<td>Care occasions/</td>
</tr>
<tr>
<td></td>
<td>(0=not readmitted, 1= readmitted)</td>
<td></td>
<td></td>
<td></td>
<td>Own calculation</td>
</tr>
<tr>
<td><strong>Comorbidity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChronicalDis</td>
<td>factor</td>
<td></td>
<td></td>
<td>28 types of chronical diseases</td>
<td>Diagnosis/</td>
</tr>
<tr>
<td></td>
<td>(28 levels)</td>
<td></td>
<td></td>
<td></td>
<td>Own calculation</td>
</tr>
<tr>
<td><strong>Miscellaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Department</td>
<td>factor</td>
<td></td>
<td></td>
<td>Hospital department where the patient was treated</td>
<td>Care occasions</td>
</tr>
<tr>
<td></td>
<td>(23 levels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TreatDem</td>
<td>numerical</td>
<td>0.83</td>
<td>0.95</td>
<td>Level of financial demandingness of the treatment</td>
<td>Care occasions</td>
</tr>
<tr>
<td><strong>DischType</strong></td>
<td>factor</td>
<td></td>
<td></td>
<td>Type of patient’s discharge location</td>
<td>Care occasions</td>
</tr>
<tr>
<td></td>
<td>(6 levels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.4.1. Logistic regression

Logistic regression models are the most widely used models in biostatistics. It is very widely used for modelling dichotomous response variables. Let the i:th observation (i = 1,...,n) in the data has response $y_i$ (yes/no), a k dimensional covariate matrix $X_i$ is associated with it, and the observations are independent. Then the logistic model is given as

$$P[y_i = 1|x_i] = \frac{e^{(a + x_i'\beta)}}{1 + e^{(a + x_i'\beta)}} \quad (1a)$$

Where $\beta$ is k dimensional parameter matrix and

$$\ln \left( \frac{p_i}{1-p_i} \right) = a + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \quad (1b)$$

The logistic regression model has many similarities with the linear regression model. The independent variables can be both continuous and categorical variables. One important advantage of the model is it can produce a simple probabilistic formula for the classification problems.

It is observed from the literature that for an unbalanced dataset, traditional classifier models may be biased towards majority class which produces poor prediction accuracy over the minority class. The data used in this study is an example of unbalanced data set, where the observations who are readmitted are 15% of the total number of observations. Before fitting any logistic model, the data were balanced in a way that all the observations with readmission were kept. Then a random sample of the same size as readmission portion was taken from the available records on the patients who were not readmitted. Then these two parts of observations were added to make a balanced data set to work on. 70% of the balanced data set was taken as the training data set based on suggestion from literature review.

Although the data have been sampled retrospectively, due to balancing, the logistic model continues to apply with same coefficients $\beta$ but a different intercept, proposed by McCullagh and Nelder (1989). However, the intercept term must be corrected for bias based on sampling proportions from different exposure groups. To explain the correction process a dummy variable $Z$ is introduced to define whether an individual is sampled or not, and denote the sampling proportions by,

$$\pi_0 = P[Z = 1|y_i = 1]$$

$$\pi_1 = P[Z = 1|y_i = 0]$$

It is obvious that the sampling proportions depend only on the occurrence of readmission not on explanatory variables. Bayes’ theorem can now be employed to compute the disease frequency among the sampled individuals who have a specified covariate vector $X$. 

18
\[ P[Y_t = 1|Z = 1, X] = \frac{P[Z = 1|Y_t = 1, X] P[Y_t = 1|X]}{P[Z = 1|Y_t = 1, X] P[Y_t = 1|X] + P[Z = 1|Y_t = 0, X] P[Y_t = 0|X]} \]

\[ = \frac{\pi_0 e^{(a + x_i' \beta)}}{\pi_1 + \pi_0 e^{(a + x_i' \beta)}} \]

\[ = \frac{e^{(a^* + x_i' \beta)}}{1 + e^{(a^* + x_i' \beta)}} \]

\[ = \frac{e^{(a^* + x_i' \beta)}}{1 + e^{(a^* + x_i' \beta)}} \quad (1c) \]

Where, \( a^* = a + \log(\pi_0 / \pi_1) \).

The logit model in this study is built in backward elimination, (see McCullagh & Nelder, 1989, pp. 89-93). The final model is achieved by eliminating insignificant variables one by one. The model is built using the global P-value of 0.20 and it is corrected for the number of variables in the model, which is 37. The simplest and most conservative approach of this correction is the Bonferroni correction, (see Banks et al., 2005) which sets the P-value for the entire set of n explanatory variables equal to P value by taking the P for each variable equal to P-value/n. The decision of keeping one variable in the model is taken based on that threshold p-value. In this study, the threshold p-value after Bonferroni correction is 0.0054.

**3.4.2. Random forests**

Random forests (RF) is a machine learning classifier. The idea behind random forest is that a number of decision trees are built using data from a training set and after training, trees are combined into one single prediction. Each of the decision trees is built separately on a bootstrapped random sample taken with replacement from the training data, and for each of the decision trees, randomly selected predictors are used. Thus randomness appears in the process twice.

In order to build the risk prediction model of hospital readmission, the final dataset consisting of 74,492 observations is divided into a training set (70% of the observations) and testing set (30% of the observations). The dependent variable in the model is NextReadm, a binary variable equal to 1 given a patient was readmitted within a 30 days period following the hospital discharge, 0 otherwise. As explanatory variables (predictors), all the available variables from the dataset, presented in Table 3 are considered. As explained in the previous part of the chapter 3, the final dataset is unbalanced with respect to the dependent variable, the share or the NextReadm in the dataset is approximately 15%.
Therefore, the training dataset for the model building is balanced with respect to the dependent variable.

Although there is little need for tuning in order to improve the classification performance of the models based on RF, there are two parameters that can be tuned. These are number of trees, \( T \), to be grown in the forest, and number of randomly assigned predictor variables at each tree, \( m \).

In order to decide the first parameter for the risk prediction model, number of trees \( T \) to be grown in the forest, a plot showing the relationship between the number of trees and so called out-of-bag error is used. This plot is presented in Figure 3. The out-of-bag error is calculated in following manner: As each tree is constructed using a different random sample with replacement, having the same sample size as the original training set, a certain number of observations of the original training set is left out and not included in the sample. These observations are called out-of-bag data and can be used to estimate the generalization error and assess variable importance, as explained later. In order to calculate the out-of-bag error, after each tree is built, the observations that didn’t make it to the randomly generated sample are used as testing data for this particular tree and the error rate is calculated. After adding each additional tree into the forest, the value of the out-of-bag error rate is averaged over all the trees built so far.

![Figure 3: Random Forests Out-of-bag error](image)

Figure 3 shows how the value of the out-of-bag (OOB) error decrease with increasing number of built trees in the training. Flattened part of the curve indicates that the number of built trees is high enough. The black curve in the Figure 3 represents the overall error, while the green and red curves represent the error connected with prediction of the particular value of the dependent variable – 1 and 0 respectively. From Figure 3 it can be observed that from a threshold of approximately 200 forests all three curves planes out, which indicate that there is no further increase of accuracy with growing
number of built trees. Based on Figure 3, it is expected that a model built on 500 trees should reach the maximal possible accuracy.

In order to decide the second parameter in the risk prediction model, \( m \) number of variables in the tree, a 10-fold cross-validation is performed. In the 10-fold cross-validation, the dataset was divided into ten subsets, where each subset was used as testing data and the remaining subsets are used as training data, for increasing number of variables from 1 to 15. This analysis has shown that the performance of the model does not substantially differ for different number of variables. Therefore, the generally recommended value of square root of the total number of available variables is used in the model.

After deciding the values for the parameter \( T = 500 \) trees and \( m = 6 \) predictors, the final model can be created and applied on the unbalanced testing dataset. For each observation in the testing dataset, predicted outcome of the dependent variable NextReadm is generated. Comparing the predicted outcome to the observed outcome for each observation, four different combinations can be obtained:

- True positive (TP): correctly classified as positive
- True negative (TN): correctly classified as negative
- False positive (FP): incorrectly classified as positive
- False negative (FN): incorrectly classified as negative

The model in this paper is built in order to predict NextReadm, i.e. whether the patient is readmitted within following 30 days or not. Thus, the true positive outcome (TP) means correctly predicted patients being readmitted. Similarly, true negative outcome (TN) means correctly predicted patients not being readmitted, false positive outcome (FP) means that a patient is being predicted to be readmitted, while in the reality no readmission happened; this value is might also be called type I error. On contrary, false negative outcome (FN) means that a patient was predicted not to be readmitted, while in the reality the readmission happened, i.e. FN represents type II error.

Summarizing number of TP, TN, FP and FN for each observation in the entire testing dataset provides information to evaluate the predicting ability of the model. These values can be inserted into confusion matrix and sensitivity and specificity might be calculated. Sensitivity/true positive rate indicates the ability of the model to identify positive patterns, i.e. it is the proportion of all patients being readmitted within 30 days following the discharge who were correctly identified by the model:

\[
Sensitivity = (TPR) = \frac{TP}{TP + FN}
\]

Specificity/true negative rate describes the ability of model to identify negative patterns; it means proportion of all patients not being readmitted within 30 days following the discharge, who were correctly predicted by the model:
Specificity \(= (TNR) = \frac{TN}{FP + TN}\)

ROC curve or c-statistics combines these two metrics, sensitivity and specificity into one single metric. The sensitivity is plotted along the y-axis, while 1-specificity is plotted along the x-axis. As shown in Tan, Steinbach, and Kumar (2014) a model with good classification performance should be located close to the upper left corner of the diagram, while model producing random guesses is located along the diagonal connecting the points TRP=0, FPR=0 and TRP=1, FPR=1. Similarly, the area under the ROC curve (AUC), yielding one single value, can be calculated to assess predictive power of the model. AUC should be close to 1 for a model with good prediction ability, and around the value 0.5 given the prediction is solely on random.

Lastly, one of the outcomes of the RF model is information on importance of the variables/predictors in the model. To this end, mean decrease of accuracy (MDA) is calculated for each predictor in the model. To assess the importance of the variables for the model, each variable is passed down in the tree and prediction is calculated. Thereafter, the values of the particular variable are randomly permuted in the out-of-bag data and in the same time, values of all other variables kept. The modified, out-of-the bag data are then passed down the tree and new predictions are computed. With help of these two datasets, the original one and the one with permuted variable, the difference of MDA is obtained. The higher the value, the more important the variable is for the model.
4. Results and Discussion

4.1. Logistic regression

Logistic model is a simple and effective tool which acts as a classifier and also specifies the factors along with their impacts on the classification process. In this study, the developed logit model can predict the readmission as well as specifies which factors are responsible for occurrence of readmission and should be taken care of. The final model was derived by backward elimination process and then the intercept term was corrected according to the sampling scheme. Table 4 shows the estimated coefficients in accordance with their magnitude, in absolute scale, as well as their corresponding standard errors and P-values.

Table 4: Coefficients estimation from the model with their standard errors and P-values

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>β coefficients</th>
<th>Standard errors</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (After correction)</td>
<td>-1.26</td>
<td>0.14</td>
<td>8.27e-04 ***</td>
</tr>
<tr>
<td>DischargeContext (to special residence)</td>
<td>-1.53</td>
<td>0.15</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>DischargeContext (to home)</td>
<td>-1.44</td>
<td>0.14</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>MetastaticCancer</td>
<td>0.68</td>
<td>0.08</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>Readmission</td>
<td>0.58</td>
<td>0.05</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>Phychoses</td>
<td>0.46</td>
<td>0.08</td>
<td>3.07e-09 ***</td>
</tr>
<tr>
<td>LiverDisease</td>
<td>0.44</td>
<td>0.08</td>
<td>2.86e-08 ***</td>
</tr>
<tr>
<td>FluidElectrolyteDisorder</td>
<td>0.43</td>
<td>0.06</td>
<td>1.79e-12 ***</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>0.34</td>
<td>0.12</td>
<td>3.20e-03 **</td>
</tr>
<tr>
<td>PepticUlcer</td>
<td>0.34</td>
<td>0.08</td>
<td>1.29e-05 ***</td>
</tr>
<tr>
<td>Coagulopathy</td>
<td>0.29</td>
<td>0.10</td>
<td>3.04e-03 **</td>
</tr>
<tr>
<td>Prev365daysReadmission</td>
<td>0.29</td>
<td>0.02</td>
<td>1.14e-10 ****</td>
</tr>
<tr>
<td>ValvularDisease</td>
<td>0.24</td>
<td>0.04</td>
<td>8.45e-08 ***</td>
</tr>
<tr>
<td>RenalFailure</td>
<td>0.22</td>
<td>0.05</td>
<td>4.89e-06 ***</td>
</tr>
<tr>
<td>Anemia</td>
<td>0.17</td>
<td>0.05</td>
<td>1.18e-03 **</td>
</tr>
<tr>
<td>HypertensionWComp</td>
<td>0.17</td>
<td>0.06</td>
<td>2.64e-03 **</td>
</tr>
<tr>
<td>OtherNeurologicalDisorder</td>
<td>0.17</td>
<td>0.05</td>
<td>2.15e-04 ***</td>
</tr>
<tr>
<td>ChronicPulmonaryDisease</td>
<td>0.15</td>
<td>0.04</td>
<td>2.56e-04 ***</td>
</tr>
<tr>
<td>Depression</td>
<td>0.13</td>
<td>0.04</td>
<td>6.34e-04 ***</td>
</tr>
<tr>
<td>Age</td>
<td>0.01</td>
<td>0.01</td>
<td>1.14e-10 ***</td>
</tr>
</tbody>
</table>

The residual deviance : 20086 (at 15806 degrees of freedom),
Pearson’s Chi-square statistics: Chi-square: 818.69 on degrees of freedom:15806 , p-value: 2.119698e-109 ; *** for significant at 0.001, ** for significant at 0.005

The results can be interpreted in following way, suppose there is a 55 years old patient who is suffering from Metastatic Cancer, his previous admission was a readmission, he had been admitted to
the hospital 3 times in last 365 days, and he is also suffering from depression. Then, the probability of his next admission to be a readmission can be calculated as follows,

The linear predictor is: \(-1.261827 + 0.68 + 0.583 + (0.297 \times 3) + 0.13 + (0.0052 \times 55) = 1.30817\)

Giving the probability of readmission = \(\exp[1.30817]/(1+\exp[1.30817]) = 0.79\)

So, according to the model, there is 79% chance that this patient will be readmitted within the next 30 days.

The most important variable according to LR model is the discharge context, to where the patient is sent after discharge. Then the second most important variable is metastatic cancer. A person suffering from metastatic cancer has 97% higher odds of being readmitted within 30 days, compared to a similar individual but not suffering from metastatic cancer. Again a person whose previous admission was a readmission has 78% higher odds that he/she will be readmitted within 30 days.

In order to assess the appropriateness of the model, a very traditional and yet effective tool is the so-called normal Q-Q plot of the Pearson’s residual and it is prepared for the final logistic model (see Figure 4). The departure of the sample quantiles of the Pearson’s residuals from normality, represented by the straight line in Figure 4, at the both ends indicate that the model did not fit very well with the data. However, it was not possible to find a better fitting model within the scope of the available data and under the assumption of linearity, in logit scale.

![Figure 4: Normal Q-Q plot of Pearson’s residual of the LR model](image)

Again AUC was used as a performance measure and it shows a good classifying capability of the model: for the final logistic model the AUC was turned out to be 0.69. Figure 5 shows the Receiver Operating Characteristic (ROC) curve for the logistic model.
The final model is applied on the test data set which resembles the true ratio of readmission in the data set and a confusion matrix is built, in order to further assess the prediction ability of the LR model. On the main diagonal in Table 5 there are correctly predicted values – true positive (TP) and true negative (TN), while on the secondary diagonal, there are false positive (FP) and false negative (FN) values. As it can be seen from Table 5 the model successfully predicts 2,648 out of 3,392 readmission, in the test data, which means that the sensitivity of the model is 78.1%. The specificity of the LR model is 52.0% as it successfully predicted 10,022 out of 19,290 patients who were not readmitted, in the test data.

Table 5: Confusion matrix for LR model for unbalanced test data set

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>C-statistics</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
<td>NextReadm = 0</td>
<td>NextReadm = 1</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td>10,022</td>
<td>9,268</td>
</tr>
<tr>
<td>NextReadm = 0</td>
<td></td>
<td>744</td>
<td>2,648</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.69</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>52.0</td>
</tr>
</tbody>
</table>

We can specify some of the probable factor that can improve performance of the model. They can be listed as follows:

- The LR model is general for all kinds of patients and diseases; however different prediction models for patients with different criteria might result better performance.
- Lack of availability of variables influencing the quality of prediction of the models which are found important in other literatures.

4.2. Random forests

A random forest (RF) model, with the dependent variable NextReadm, and explanatory variables presented in Table 3, was developed. As mentioned previously, the entire dataset available for the
modelling consisted of 74,492 observations. This dataset was then divided into two parts, training set (70%) and testing set (30%). The training set consists of 52,144 observations, out of which 8,759 readmissions, the testing set consists of 22,348 observations, with 3,366 readmissions. For the purpose of training of the RF model, the training dataset was balanced in order to not to prioritize any of the predicted outcomes.

In order to evaluate the discriminating ability of the final RF model, a confusion matrix is presented in Table 6. On the main diagonal in this table there are correctly predicted values – true positive (TP) and true negative (TN), while on the secondary diagonal, there are false positive (FP) and false negative (FN) values. Table 6 shows that the model correctly predicted 2,183 out of 3,366 readmissions, which means that the sensitivity of the model is 64.9%. The specificity of the RF model is 69.5 % as it correctly predicted 13,198 patients that were not readmitted out of 18,982.

Table 6: Confusion matrix of the Random Forests model

<table>
<thead>
<tr>
<th>Observed NextReadm = 0</th>
<th>NextReadm = 1</th>
<th>C-statistics</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NextReadm = 0</td>
<td>13,198</td>
<td>5,784</td>
<td>0.60</td>
<td>64.9</td>
</tr>
<tr>
<td>NextReadm = 1</td>
<td>1,183</td>
<td>2,183</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To assess the importance of the explanatory variables in the model for prediction of hospital readmission, Figure 6 presents a plot with mean decrease of accuracy (MDA) for those explanatory variables with MDA above the value of 30. This plot is ordered top-to-bottom from most to least important variables. The most influential variable for prediction of hospital readmissions is Diagn, i.e. explanatory variable which describes the main diagnosis related to the current admission of the patient. Another two influential explanatory variables are Deparment, hospital department at which the patient was treated and PrevReadm, and the number of readmissions during the previous calendar year. There are additionally four variables, which might be perceived as influential for the prediction of patients’ readmissions, are: TreatDem, Readmission, Municipality and LenghtOfStay. The remaining three variables outlined in the Figure 6 are: Severity, Transfer, and Age, i.e. severity of the current diagnosis, type of transfer from the hospital and patient’s age, contribute relatively poorly to the prediction of the hospital readmissions. However, it is important to keep in mind that Figure 6 illustrates variable’s importance for variables giving the value of MDA equal to or greater than 30. Additional 26 variables, with mean square accuracy in the range between 8 and 30 are left out from the plot.
Figure 6: Variable accuracy in RF model

Figure 6 indicates that the most influential variable for prediction of hospital readmission is variable Diagn, i.e. explanatory variable which describes the main diagnosis related to the current admission of the patient. In other words, knowing for which disease or disorder the patient is currently treated has impact upon the outcome of the RF model. This result seems to be in line with several of the reviewed studies (e.g. Edgren et al., 2012; Krumholz et al., 2011), discussed in Chapter 2. Contrary to the patient’s current diagnosis, another variable describing the overall health condition of the patient, severity of the diagnosis, is found to have relatively low impact upon the prediction of next readmission.

Another interesting finding revealed by Figure 6 is that mainly the variables describing hospital admission history along with previous use of medical services are considered as influential variables in the RF model for the prediction of hospital readmissions. PrevReadm, Readmission and LenghtOfStay are all among the explanatory variables with mean decrease accuracy above 30. The relatively most important one among these three variables is the variable PrevReadm, ranked on the third placed in the Figure 6. In other words, knowing the patient’s history with respect to how many times he was already readmitted to the hospital within the past 365 days (PrevReadm) contributes to the correct prediction of patient’s next readmission. Similarly, also the fact that patient’s current hospital stay is already a readmission, which is identified by the variable Readmission, contributes also to the prediction of patient’s next readmission. These findings are in line with results reported in the papers reviewed in Chapter 2. Bottle et al. (2006) in the study of readmission to the emergency department of UK hospitals reports: “… strong relation between high-impact users and previous admission history. Nearly half of the patients who had three and more emergency admissions in previous year went on readmission”. Hasan et al. (2010) in a study developing a prediction model for general medicine patients reports that patients with more than one hospital admission in the last year run significantly
higher risk of hospital readmission, and furthermore, that length of stay of the current hospital spell is an important predictor of hospital readmission as well. Also the only Swedish study, Edgren et al. (2012) found that both the number of previous readmissions as well as the fact whether the current visit is a readmission are important predictors of readmission. The strongest relationship was found for having 5 or more readmission within the last year.

It is interesting to note, that several variables, pointed out as statistically significant predictors of hospital readmission in several of the reviewed studies, such as presence of chronic diseases (e.g. Hasan et al., 2010; van Walraven et al., 2010) or age (e.g. Coleman et al., 2004; Halfon et al., 2006; Silverstein et al., 2008) were not found as influential predictors in this model.

In order to measure the discriminating ability of the model, the area under the ROC curve (AUC) metrics was used. Discrimination ability of the RF model is only modest: the AUC value is 0.60. Although this value can be perceived as rather low, it corresponds to the lower band of several of the reviewed studies in Table 1 (e.g. Hammill et al., 2011; Hasan et al., 2010; Krumholz et al., 2011; Lindenauer et al., 2011). Figure 7 shows the ROC curve of the RF model.

![ROC curve of the Random Forests model trained with balanced data](image)

Figure 7: ROC curve of the Random Forests model trained with balanced data

The results presented so far are related to RF model trained on a dataset that is balanced with respect to the dependent variable NextReadm. In the context of readmission predictions, FP and FN might be related to certain “costs”. In simplicity, costs related to FP can be costs related to the treatment of patients, that were by the model identified as running high-risk of readmission, and in reality, were not readmitted. Costs related to FN represent costs related to the readmissions of patients, for which the model incorrectly predicted a low-risk of readmission, but the patients were readmitted.
However, in reality, the costs associated with additional treatment of a patient, who would not be readmitted and the costs related to hospitalization of a patient, originally indexed as running low-risk of readmission, do not need to be equal. In certain situations, it may be for instance relatively cheaper to provide preventive treatment to a larger group of patients, rather than to pay for the medical treatment of one patient, who is in need for readmission. In contrary, there might also be situation when the preventive interventions are more costly compared to the medical treatment.

In order to illustrate the behaviour of the model in situations, where costs connected with preventive interventions versus treatment of readmitted patients are not equal, additional models with different weights on FP and FN were built. These different weights were operationalized through a change of the ratio of zeros and ones of the dependent variable – NextReadm - in the training dataset (Tan et al., 2014).

Table 7: Random forests models with different proportions of readmission in the training data

<table>
<thead>
<tr>
<th>Ratio of “0” and “1”</th>
<th>Confusion matrix*</th>
<th>C-statistics</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Geometric mean of sensitivity and specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NextReadm 5:1</td>
<td>18,806 156</td>
<td>0.71</td>
<td>7.0</td>
<td>99.1</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>3,132 234</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2:1</td>
<td>16,310 2,672</td>
<td>0.62</td>
<td>41.5</td>
<td>85.9</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>1,967 1,399</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:1</td>
<td>13,198 5,784</td>
<td>0.60</td>
<td>64.9</td>
<td>69.5</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>1,183 2,183</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:2</td>
<td>6,472 12,510</td>
<td>0.57</td>
<td>90.7</td>
<td>34.1</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>317 3,049</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:5</td>
<td>571 18,411</td>
<td>0.56</td>
<td>99.6</td>
<td>3.0</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>14 3,352</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The rows of the confusion matrices show the observed 0’s and 1’s while the columns show their predicted counts.

Table 7 presents the results of these additional models, in which the ratio of zeros/ones varies from 5:1 to 1:5. For each combination, confusion matrix, geometric means of sensitivity and specificity, along with c-statistics and sensitivity / specificity is provided. First row in Table 7 represents the situation when the cost difference between costs of prevention and costs of medical treatment associated with patients’ readmission is small. It may be observed that in this setting, the model predicts a high number of negative outcomes NextReadm. Hence, such model has high specificity (99.1%), it means that majority of the negative outcomes in the dataset are correctly predicted by the model. However, this is at the expense of sensitivity, which is very low (6.9%) and thus, only few of the observed positive outcomes are actually correctly predicted by the model. On the contrary, the last row in Table 7 represents the situation, when the cost difference between costs of prevention and costs of treatment of a patient that needs to be readmitted, are large. This model has high sensitivity (99.6%) and low
specificity (3%). In other words, the model successfully predicts majority of the positive outcomes, however, produce also relatively large number of false positive outcomes.

It is also interesting to note the behaviour of the geometric mean of sensitivity / specificity and of the c-statistics for the different settings of the model in Table 7. While the geometric mean has a maximum value, it means the optimum, when the weights are equally divided between both types of errors, the c-statistics is increasing from the value 0.56 in the model with weights on positive outcomes to a value of 0.71 in the model with weights on negative outcomes. In other words this illustrates that c-statistics prioritizes the model that predicts correctly prevalent outcome of the dependent variable.
5. Conclusion

Hospital readmission is one of the drivers of medical and healthcare expenditures. Risk prediction models might assist to identify patients running high-risk of hospital readmission, so interventions to reduce the readmission risk might be targeted. Using dataset originating from medical and administrative records of patients hospitalized in one of the hospitals and primary care centres of the Dalarna County, Sweden, and defining readmission as unplanned admission of a patient into the hospital within 30 days after the original hospitalization, this study suggested two risk prediction models.

The random forests (RF) model was able to correctly predict 2,183 readmitted patients, with sensitivity 64.8% and specificity 69.5%. This model reached discriminating ability measured with c-statistics equal to 0.60. The model based on logistic regression (LR) was able to correctly predict 2,648 readmitted patients, with sensitivity 78% and specificity 52%. The c-statistics of the LR model was equal to 0.69 while tested without corrected intercept term and on balanced test data. Based on the c-statistics and sensitivity, the LR model performed slightly better. However, this is at the cost of low specificity, in other words poor prediction of the prevalent outcome in data, readmission. Thus, each of these models has its strengths and shortcomings.

In addition to the differences in the predictive power of the two models, it may be seen that the RF model performs well in situations when a pure classifier is needed. In the same time, it does not provide information on the exact effect of each of the variables included in the model on the risk of hospital readmissions. On the other hand this information can be easily retrieved from the LR model. Thus, the choice of the method for readmission risk modelling should be informed by the specific needs.

Even though the overall predictive power of both LR and RF models could be perceived as rather poor, the suggested models might be used to identify high-risk patients for whom post-hospital interventions are appropriate. This can increase cost-effectiveness of the medical care and improve quality of life of the patients, which might be appealable in situation when post-hospital care is costly and needs to be applied selectively. Furthermore, the suggested prediction models were developed using administrative data, which are assumed to be readily available in the time when patient’s discharge is planned. Hence, both the LR and RF models should provide relevant information on patient’s risk incidence already in the discharge time and in this way should assist in planning the post-hospital interventions such as follow-up visits to physicians.

There are some limitations of this study. Firstly, the definition of readmission implemented in this study is all acute readmissions of the patients within 30 days after the initial hospitalization. No distinction is done between readmissions associated with the initial diseases or disorder, and
readmission for a new condition. Although it is not expected that the latter type of readmission is frequently occurring, this might overestimate the prediction results. Secondly, the study is limited by its reliance on the previously mentioned administrative data, therefore, several variables deemed important in some published prediction models for hospital readmission could not be replicated. Thirdly, although this study provides evidence on factors being associated with readmission risks for all patients treated by the hospitals in Dalarna County, it does not bring much clarity into factors affecting the readmission risk for group of patients with specific diagnosis.

Further research could, instead of modelling readmission risk for a general population, as was done in our study, focus on modelling readmission risks of a specific subpopulation. Preferably those patients that stands for a large share of the hospital readmissions, and thus, for a large share of costs. Patients with heart failure diagnosis and surgical patients are two commonly used subpopulations, as was found in the literature review. Elderly patients are another subpopulation, for which the risk prediction models could be done individually.

Another interesting venue for the future research would be to incorporate into the model clinical data, i.e. data related to various medicine tests taken in connection to the current hospitalisation. Clinical data have been proven by several studies as contributing to higher prediction ability of the models. This would, however, require a careful consideration on which of the clinical data can be obtained in relatively short time period following the admission, so the risk predictions could be calculated also for patients with relatively short hospitalisation time.

Lastly, as mentioned earlier, hospital readmissions are mainly associated with unnecessary costs, and readmission modelling is adopted in order to identify patients running higher risks of readmission to be able to target risk-reducing interventions. However, none of the models in our study can fully identify the risks for each of the patients, and thus, besides correctly identified outcomes for readmitted/not readmitted patients; a certain number of outcomes are identified incorrectly. These incorrectly identified outcomes are also related to certain costs. It would therefore be an interesting issue for further research, to incorporate weights into the prediction models, and in this way, accommodate the model for different costs related to incorrectly identified outcomes.
References


