CHAPTER 3

Identifying patients at increased risk of hypoglycaemia in primary care: Development of a machine learning-based screening tool

Stijn Crutzen, Sunil Belur Nagaraj, Katja Taxis and Petra Denig

Diabetes/Metabolism Research and Reviews (2020), (November).
CHAPTER 3

Abstract

Introduction
In primary care, identifying patients with type 2 diabetes (T2D) who are at increased risk of hypoglycaemia is important for the prevention of hypoglycaemic events. We aimed to develop a screening tool based on machine learning to identify such patients using routinely available demographic and medication data.

Methods
We used a cohort study design and the Groningen Initiative to ANalyse Type 2 diabetes Treatment (GIANTT) medical record database to develop models for hypoglycaemia risk. The first hypoglycaemic event in the observation period (2007-2013) was the outcome. Demographic and medication data were used as predictor variables to train machine learning models. The performance of the models was compared with a model using additional clinical data using 5-fold cross validation with the area under the receiver operator characteristic curve (AUC) as a metric.

Results
We included 13,876 T2D patients. The best performing model including only demographic and medication data was logistic regression with least absolute shrinkage and selection operator, with an AUC of 0.71. Ten variables were included (odds ratio): male gender (0.997), age (0.990), total drug count (1.012), glucose-lowering drug count (1.039), sulfonylurea use (1.62), insulin use (1.769), premixed insulin use (1.109), insulin count (1.827), insulin duration (1.193), and antidepressant use (1.05). The proposed model obtained a similar performance to the model using additional clinical data.

Conclusion
Using demographic and medication data, a model for identifying patients at increased risk of hypoglycaemia was developed using machine learning. This model can be used as a tool in primary care to screen for patients with T2D who may need additional attention to prevent or reduce hypoglycaemic events.
Identifying patients at increased risk of hypoglycaemia in primary care

**Introduction**

Strict glucose control in patients with type 2 diabetes (T2D) is essential for preventing long-term micro- and macro-vascular complications, such as retinopathy, neuropathy, and cerebrovascular, kidney and heart disease. However, strict control of T2D with insulin and insulin secretagogues, especially in frail and older patients, increases the risk of hypoglycaemia. Commonly, three levels of hypoglycaemia are distinguished. Glucose levels between 3.9 mmol/L and 3.0 mmol/L are defined as mild hypoglycaemia and are considered an alert value. Glucose levels below 3.0 mmol/L are defined as moderate hypoglycaemia and are considered clinically important events, whereas severe hypoglycaemia is defined by a mental or physical impairment that requires external assistance. Hypoglycaemia is not only associated with reduced quality of life and higher healthcare costs but also hinders strict glucose control. Severe cases of hypoglycaemia can lead to unconsciousness, hospitalization, and even brain death. Hypoglycaemia during hospital admissions has been associated with prolonged hospital stay and increased mortality. In addition to the mortality directly caused by hypoglycaemic events, several studies have shown an association of hypoglycaemia with all-cause and cardiovascular mortality.

In T2D patients, the risk of hypoglycaemia strongly varies among patients. Not all T2D patients require insulin or insulin secretagogues to control their glucose level. Moreover, differences in co-medication can influence the risk of hypoglycaemia. For instance, non-selective beta-blockers reduce the effects of epinephrine, which in turn, increases hypoglycaemia unawareness, and selective serotonin reuptake inhibitors can increase sensitivity to insulin. In addition to differences in prescribed medications, differences in comorbidities can also influence the risk of hypoglycaemia. Dementia, for instance, reduces patients’ ability to manage their glucose-lowering medications and increases the risk of medication errors, and depression can lead to poor self-care and self-monitoring. Furthermore, decreased renal and liver functions can diminish the clearance of both endogenous and exogenous insulin.

Several models have been developed to identify patients at high risk of hypoglycaemia. Most of these models have been developed using a mix of type 1 diabetes patients and T2D patients, focus on selecting high risk patients in inpatient settings, and often include previous hypoglycaemic events to predict future events. Most hypoglycaemic events occur outside the hospital setting and therefore need to be managed by primary care. In primary care, stratifying patients based on hypoglycaemia risk can provide an approach in which treatment, glycaemic goals and education are tailored to the individual, which ultimately helps to reduce hypoglycaemia. Previous research has shown that pharmacy-led interventions can be effective in reducing the number of hypoglycaemic events in T2D patients. Using routinely available pharmacy data to screen for patients at...
increased risk of hypoglycaemia can improve the effectiveness and efficiency of these interventions. Generally, pharmacy data include information about patients’ medication history and all drugs dispensed during a specific period. Conventional approaches to using such information in risk models include creating simple measures of medication use at the substance or therapeutic class level. Recent developments in machine learning provide statistical tools to include multiple characterizations of medication use in one model, for example, ‘current insulin use’, ‘insulin count’, ‘insulin type’ and ‘insulin use duration.’ In this way, more information can be utilized for risk modelling.

In this study, we aimed to develop a screening tool to identify T2D patients at increased risk of hypoglycaemic events based on machine learning using demographic and medication data available in community pharmacies. Additionally, we investigated whether the performance of the model could be improved by adding electronic clinical data that are available in general practice. We also compared the results of these models with results from traditional regression models.

Material and methods
Study design
We conducted a cohort study using demographic and medication data to predict the occurrence of hypoglycaemia in T2D patients. The performance of this model was compared to that of a model that additionally included clinical data. Data from the Groningen Initiative to ANalyse Type 2 Diabetes Treatment (GIANTT) were used (www.giantt.nl). This database includes data from a longitudinal dynamic cohort of more than 60,000 patients with a GP-confirmed diagnosis of T2D. The data were extracted from the electronic medical records of 180 general practices (GPs) in the northern parts of the Netherlands. At the end of our study period, approximately 80% of all the GPs in this region contributed to the GIANTT database. The database contains patients’ demographic data, data on prescribed medications (similar to medication data in a pharmacy information system), medical history based on the International Classification of Primary Care (ICPC) and routinely available clinical measurements, such as blood pressure and haemoglobin A1c (HbA1c) levels.

Outcome and predictor variables
The outcome of this study was the first hypoglycaemic event during the observation period between January 2007 and January 2014. Hypoglycaemic events were defined by the ICPC code for hypoglycaemia (T87) or by a recorded glucose measurement below 3.9 mmol/L. We therefore included all the levels of hypoglycaemia that were recorded in the medical records by GPs since mild events can also be relevant for our screening tool. Recorded hypoglycaemic events in free
text parts of the medical record were used to enrich the T87 codes in the GIANTT database. The date of the first hypoglycaemic event was used as the index date. Patients without any hypoglycaemic events were assigned a random index date in the observation period. When assigning these random dates, the distribution of the index dates of the event group was taken into account to correct for temporal influences. Based on the literature, potential predictor variables related to demographics, medication and clinical information were selected. An overview of the selected variables and their definitions can be found in online appendix I. Any prescription in the four months prior to the index date was used to determine current medication use, the latest measured value within one year prior to the index date was used for most clinical measurements and active ICPC codes at the time of the index date were used for comorbidities. For the medication data, potential predictor variables included not only the current use of medication but also other aspects of medication use, including duration of use, counts of medication classes, and potential drug-drug interactions.

**Study population**
This study included T2D patients who were 18 years or older and were treated for diabetes in a primary care setting. We excluded patients registered with a GP with poor documentation of hypoglycaemia, as demonstrated by a lack of recordings for hypoglycaemia (T87) for any of their T2D patients. The following patient exclusion criteria were applied: A) included in the database for less than one year prior to their index date; B) no prescription of a glucose-lowering medication one year prior to the index date; C) no T2D diagnosis before or within six months after the index date; and D) missing all of the following measurements one year prior to the index date: estimated globular filtration rate (eGFR), creatinine, albumin to creatinine ratio (ACR), HbA1c, weight, body mass index (BMI), high density lipoprotein (HDL), low density lipoprotein (LDL), total cholesterol, systolic blood pressure (SBP) and diastolic blood pressure (DBP).

Due to the imbalanced dataset (N=2,523 and 11,344 with and without a hypoglycaemic event, respectively), which could severely bias the performance of machine learning (ML) algorithms, we used a sample-size equalization method to balance the data. In this method, we randomly divided the patients without a hypoglycaemic event into four equal subdatasets, corresponding to the number of patients in the hypoglycaemic group, as shown in figure 1 (11,344/2,523=4 subdatasets). Model training was performed using these balanced datasets.
Figure 1. Schematic overview of the sample-size equalization method and fivefold cross-validation to evaluate and compare the performance of the different machine learning models. To balance the data, the non-hypoglycaemia patients were divided in four equal groups, which were each matched with the hypoglycaemia patients. This was followed by fivefold cross-validation in each of the four subdatasets to determine which machine learning method in which subdataset resulted in the best performing model using area under the curve (AUC) as metric.
Analysis
We first evaluated the prediction performance of all the individual variables and evaluated the performance of traditional logistic regression using either all the demographic and medication data or all the demographic, medication and clinical data. Next, to develop the screening tool, we trained several machine learning algorithms on either demographic and medication data or demographic, medication and clinical data. We evaluated the performance of the following algorithms: logistic regression with backwards selection, ridge logistic regression, elastic net logistic regression ($\alpha = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ or $0.9$), least absolute shrinkage and selection operator (LASSO) logistic regression and random forest (RF). These algorithms were used because they are relatively simple to interpret, easy to implement in practice and, in the cases of elastic net, LASSO and RF, are suitable for dealing with multidimensional data.\textsuperscript{39,40} Missing clinical measurement data were imputed using the K-nearest neighbour ($k=5$) multiple imputation method (10-fold). The list of the predictor variables and the percentage missing for each variable are given in online appendix I. We performed 5-fold cross-validation (80% training data, 20% testing data) to evaluate the performance of the models using AUC as a metric. The best performing model without clinical data was compared with the best performing model including clinical data. Stata 15.0 was used for data cleaning, logistic regression and RF. R version 3.5.1 was used with the glmnet package for LASSO, elastic net, and ridge logistic regression.\textsuperscript{41} For RF, out of the bag error was used to find the optimal number of trees and the optimal number of variables considered at each node. No maximum depth was set, and the minimum number of observations per leaf was set to one. We performed 5-fold cross-validation to identify optimal values for the lambda plus one standard error for LASSO, elastic net, and ridge logistic regression. Second, we performed 5-fold cross-validation to identify the optimal $\alpha$ for the elastic net between 0.1 and 0.9 with steps of 0.1. Third, we performed 5-fold cross-validation to select the balanced subdataset and the imputed dataset that resulted in the best performance. The predictor variables were standardized for the elastic net, ridge and LASSO logistic regression analyses. A heatmap was created to show the importance of the variables used in the best performing model.

Ethics statement
Based on the research code of conduct in the Netherlands, research using anonymous medical record data requires no ethics committee approval.
CHAPTER 3

Results

We extracted 40,124 patients from the GIANTT database. After exclusion, 13,876 patients remained, 19.2% of whom had at least one hypoglycaemic event during the observation period (online appendix II). The subdatasets contained either 5,201 or 5,202 patients. Table 1 shows that patients with and without a hypoglycaemic event were similar in age and clinical measurements, but patients with an event used insulin more often and had a longer diabetes duration.

Table 1. Characteristics of patients with and without a hypoglycaemia event

<table>
<thead>
<tr>
<th>Variables</th>
<th>Hypoglycaemia patients</th>
<th>No hypoglycaemia patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of patients, n (%)</td>
<td>2,523 (19.1)</td>
<td>10,713 (80.9)</td>
</tr>
<tr>
<td>Age, years</td>
<td>66.4 (12.5)</td>
<td>67.9 (12.1)</td>
</tr>
<tr>
<td>Female, %</td>
<td>45.0</td>
<td>50.2</td>
</tr>
<tr>
<td>Diabetes duration, years</td>
<td>8.3 (6.5)</td>
<td>6.8 (5.6)</td>
</tr>
<tr>
<td>Insulin use(^a), %</td>
<td>34.0</td>
<td>12.1</td>
</tr>
<tr>
<td>Sulfonylurea use(^a), %</td>
<td>44.6</td>
<td>37.1</td>
</tr>
<tr>
<td>Metformin use, %</td>
<td>75.5</td>
<td>79.0</td>
</tr>
<tr>
<td>Number of medicines</td>
<td>6.5 (3.5)</td>
<td>5.8 (3.3)</td>
</tr>
<tr>
<td>HbA1c, %</td>
<td>7.1 (1.0)</td>
<td>7.0 (1.0)</td>
</tr>
<tr>
<td>BMI, kg/m(^2)</td>
<td>30.2 (5.6)</td>
<td>30.0 (5.5)</td>
</tr>
<tr>
<td>SBP, mmHg</td>
<td>138.6 (17.0)</td>
<td>139.6 (17.7)</td>
</tr>
<tr>
<td>DBP, mmHg</td>
<td>77.3 (9.6)</td>
<td>78.1 (10.1)</td>
</tr>
<tr>
<td>LDL, mmol/L</td>
<td>2.5 (0.88)</td>
<td>2.5 (0.91)</td>
</tr>
<tr>
<td>HDL, mmol/L</td>
<td>1.2 (0.34)</td>
<td>1.2 (0.35)</td>
</tr>
<tr>
<td>Total cholesterol, mmol/L</td>
<td>4.4 (1.0)</td>
<td>4.4 (1.2)</td>
</tr>
<tr>
<td>eGFR, ml/min/173 m(^2)</td>
<td>75.1 (22.8)</td>
<td>76.7 (23.2)</td>
</tr>
</tbody>
</table>

Values are reported as the mean with the standard deviation (sd) unless reported otherwise. Abbreviations: HbA1c= haemoglobin A1c, BMI= body mass index, SBP= systolic blood pressure, DBP= diastolic blood pressure, LDL= low density lipoprotein, HDL= high density lipoprotein, eGFR= estimated glomerular filtration rate. \(^a\) Of the hypoglycaemia patients, 71.4% used insulin and/or sulfonylurea, and 46.9% of the patients without hypoglycaemia used insulin and/or sulfonylurea.

Models with demographic and medication data

Using the demographic and medication data, the best predicting individual variable was ‘insulin use duration’, with an AUC of 0.64 (figure 2). Out of the 25 individual variables, three had an AUC above 0.6, six had an AUC between 0.60 and 0.55, and the remaining 16 variables had an AUC below 0.55. Using traditional logistic regression without variable selection resulted in an AUC of 0.69. The best performing machine learning algorithm was the LASSO logistic regression algorithm, with a mean AUC of 0.71 (±0.019) using 10 variables (table 2). The LASSO
Identifying patients at increased risk of hypoglycaemia in primary care

logistic regression algorithm selected age, sex, six diabetes medication-related variables and two co-medication-related variables. The most important variable was ‘insulin use’, followed by ‘sulfonylurea use’ and ‘insulin use duration’ (figure 3). Although sex and age were selected as predictors in most folds, their importance was relatively low. Logistic regression with backward selection resulted in a mean AUC of 0.69, ridge logistic regression resulted in an AUC of 0.69, the best performing elastic net logistic regression model ($\alpha = 0.6$) resulted in an AUC of 0.70 and RF resulted in an AUC of 0.68.

![Figure 2. Boxplot of the area under the curve of the individual predictors, based on 1000 bootstraps. AH, antihypertensive; BB, beta-blocker; cortico., corticosteroid; chemo., antineoplastic or immunomodulating agent; DPP-4, dipeptidyl peptidase 4 inhibitor; GLD, glucose-lowering drugs; drug interaction, interaction between insulin and/or sulfonylurea with co-medication; ins., insulin](image-url)
Figure 3. Heatmap showing the set of variables selected in the different folds of the fivefold cross-validation in the four subdatasets (5x4 folds) based on least absolute shrinkage and selection operator (LASSO) logistic regression algorithm. A darker blue colour represents a higher weight assigned by LASSO. This is indicative of a higher importance of a variable for predicting hypoglycaemia. AH, antihypertensive; BB, beta-blocker; cortico., corticosteroid; chemo., antineoplastic or immunomodulating agent; DPP-4, dipeptidyl peptidase 4 inhibitor; drug interaction, Interaction between insulin and/or sulfonylurea with comedication; GLD, glucose-lowering drugs; ins., insulin
Models with additional clinical data
The best predictive individual clinical variable was diabetes duration, with an AUC of 0.57 (online appendix III). When including additional clinical data, traditional logistic regression without variable selection resulted in an AUC of 0.71. The best performing machine learning algorithm was the LASSO logistic regression algorithm, with an average AUC of 0.71 (±0.024). The performance of the resulting model was similar to that of the model without clinical data. The model included nine of the same demographic and medication variables; three additional medication variables: antipsychotic, antibiotic and oral corticosteroid use; and ten additional clinical variables: diabetes duration, weight, estimated glomerular filtration rate (eGFR), HbA1c, total cholesterol, depression, high blood pressure, non-chronic infection, hypercholesterolemia and albuminuria (table 2). Logistic regression with backward selection resulted in an average AUC of 0.71, and ridge logistic regression resulted in an AUC of 0.70. For the best performing elastic net logistic regression model (α=0.5), the AUC was 0.71, and RF resulted in an AUC of 0.66.
Table 2. Odds ratios of the predictor variables for the best performing models (LASSO) using only demographic and medication data or additional clinical data

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Odds Ratio</th>
<th>Predictors</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic and medication data</td>
<td></td>
<td>Additional clinical data</td>
<td></td>
</tr>
<tr>
<td>Sulfonylurea use</td>
<td>1.620</td>
<td>Sulfonylurea use</td>
<td>2.001</td>
</tr>
<tr>
<td>Insulin use</td>
<td>1.769</td>
<td>Insulin use</td>
<td>1.660</td>
</tr>
<tr>
<td>Pre-mixed insulin use</td>
<td>1.109</td>
<td>Pre-mixed insulin use</td>
<td>1.481</td>
</tr>
<tr>
<td>Insulin use duration capped at 5 years (years)</td>
<td>1.193</td>
<td>Insulin use duration capped at 5 years (years)</td>
<td>1.244</td>
</tr>
<tr>
<td>Count of different types of insulin</td>
<td>1.827</td>
<td>Count of different types of insulin</td>
<td>1.120</td>
</tr>
<tr>
<td>Count of glucose lowering drugs</td>
<td>1.039</td>
<td>Count of glucose lowering drugs</td>
<td>1.016</td>
</tr>
<tr>
<td>Count of all drugs</td>
<td>1.012</td>
<td>Count of all drugs</td>
<td>1.008</td>
</tr>
<tr>
<td>Antidepressant use</td>
<td>1.050</td>
<td>Antidepressant use</td>
<td>1.184</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.990</td>
<td>Age (years)</td>
<td>0.986</td>
</tr>
<tr>
<td>Sex (0=F/1=M)</td>
<td>0.997</td>
<td>Antibiotic use</td>
<td>1.047</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.965</td>
<td>Oral corticosteroid use</td>
<td>1.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antipsychotic use</td>
<td>0.922</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes duration capped at 5 years (years)</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weight (kilogram)</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eGFR (ml/min/1,73 m²)</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HbA1c (%)</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total cholesterol (mmol/L)</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Depression</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High blood pressure</td>
<td>1.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-chronic infection</td>
<td>1.536</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypercholesterolemia</td>
<td>0.909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Albuminuria</td>
<td>1.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intercept</td>
<td>3.196</td>
</tr>
</tbody>
</table>

Abbreviations: HbA1c= haemoglobin A1c, eGFR=estimated glomerular filtration rate, F=female, M=male

Discussion

Main findings

A model including 10 demographic and medication variables based on machine learning showed an acceptable performance to screen for an increased risk of hypoglycaemia in T2D patients in primary care. The inclusion of additional clinical data did not improve the performance of the model.
Comparison with previous research
Several models have been developed for predicting hypoglycaemia\textsuperscript{24-30}. However, these models mainly focus on the inpatient setting or do not make a distinction between type 1 and type 2 diabetes. Our model is intended for primary care using demographic and medication data that are widely available to screen for T2D patients with an increased risk of hypoglycaemia. By including only T2D patients in the development of the model, variables that are specific for T2D patients, such as sulfonylurea use, can contribute to the model.

The performance of our model, which had an AUC of 0.71, is considered acceptable\textsuperscript{42} and comparable with several previously developed models\textsuperscript{24-26}, although some models have shown higher AUCs than ours\textsuperscript{28-30,43}. The higher performance of these models may be due to the availability of richer data in clinical trials and inpatient settings in comparison to data that are routinely available in outpatient settings. For example, daily glucose measurements may be available for diabetes patients who are admitted to a hospital but not for T2D patients in primary care. More importantly, all of the better performing models included prior hypoglycaemic events as a variable, which was one of the, if not the most, important predictor in these models. Since we aimed to screen for patients at increased risk of a first hypoglycaemic event, we did not use prior events as a predictor. Predicting the first hypoglycaemic event is more difficult, but it is essential to identify at-risk patients who are not already known to healthcare professionals.

Many of the variables in our models are known risk factors for hypoglycaemia, such as using different types of insulin or using pre-mixed insulin. Previous research has shown that a longer diabetes duration is predictive of hypoglycaemia\textsuperscript{30}. In our models, as in previous research, a longer duration of insulin use was associated with a higher risk of hypoglycaemia\textsuperscript{44}. When clinical data were added to that model, diabetes duration—in addition to a longer insulin use duration—was associated with a lower risk. Surprisingly, lower age contributed to an increased risk of hypoglycaemia in our models, whereas in other models, higher age was predictive of hypoglycaemic events\textsuperscript{26,28-30}. This difference might be due to differences in the severity of the hypoglycaemia events used as outcomes. Previous studies mainly used severe hypoglycaemic events as the outcome, while we included any events regardless of severity. It has been found that younger patients may report more hypoglycaemic events, in general, but not more severe hypoglycaemic events\textsuperscript{45}. This finding might be due to the increased hypo-unawareness in older patients\textsuperscript{46}. The inclusion of antidepressant use as a predictor could be explained by the increase in insulin sensitivity caused by selective serotonin reuptake inhibitors\textsuperscript{19}. Another explanation might be the association of severe depression with severe hypoglycaemia in T2D patients\textsuperscript{22}.

Of note is our finding that clinical data, such as the HbA1c level, did not improve the performance of the model. Additionally, the HbA1c level was similar for those who did and those who did not have a hypoglycaemic event. This
finding suggests that the level of glucose control is not a good measure to inform healthcare professionals about the risk of hypoglycaemia, which is in line with previous research that showed that HbA1c is not related to hypoglycaemia risk in older T2D patients who use insulin. More generally, it is important to realize that by developing a model based on a large amount of data, some patient characteristics that are considered clinically relevant may not contribute to the performance of the model as a whole.

**Strengths and limitations**

By using machine learning to develop our models, we were able to include several related variables, which allowed us to make better use of the information available when only using medication data. Particularly, considering more information about insulin than just current use improved the performance of the model. By using LASSO to select variables, a simpler model could be obtained. LASSO logistic regression is able to select predictors without having to rely on P-values, which are highly dependent on power. Another strength is the use of cross-validation, which provides a less biased estimation of the performance of models. The use of a large database, which consists of routinely available primary care data, has two advantages. First, using routinely available data will mimic daily practice more closely, resulting in a more accurate estimation of the performance of the prediction models. Second, by using these data, only predictors that are commonly available and well documented in primary care are included in the prediction model.

Several limitations of the study should be noted. First, it is likely that a portion of the control patients were misclassified. By using routinely available data, we missed hypoglycaemic events that were not documented by GPs. Although we excluded GPs who never documented hypoglycaemic events, the included GPs may not have consistently documented such events. This misclassification most likely resulted in a lower performance of the models. Bias could have been introduced when the missed events were not random. For instance, GPs might be more primed to recognize hypoglycaemia in patients who use insulin, causing an underestimation of hypoglycaemia risk in patients who do not use insulin. Second, some patients may have been falsely classified as not using insulin or sulfonylurea because the four-month look-back period was too short for prescriptions that can last for a longer period. When this period was increased to twelve months, insulin use increased by approximately five percent points in the hypoglycaemia group. A third limitation is that we only performed an internal validation. Finally, four predictors that were selected by LASSO in the best performing model with clinical data were imputed, namely, weight, eGFR, HbA1c and total cholesterol. This limits the applicability of the model using clinical data. In addition, potential predictors not documented, for example, related to the level of education or self-management, were not included, and the use of newer diabetic medication was not included in our primary care dataset in the study period.
Identifying patients at increased risk of hypoglycaemia in primary care

**Implication for practice**

Both the models including demographic and medication data and the model with additional clinical data could be used to identify patients with type 2 diabetes in primary care who are at increased risk of hypoglycaemia. For implementation in clinical practice, we recommend a screening tool based on the model without clinical data because it requires less information and shows a similar performance to the model with clinical data. Our model can be easily implemented in the electronic information systems of community pharmacies as well as GPs, converting the ORs from the LASSO model into a risk score. The model is not intended to replace or mimic the clinical judgement of a healthcare professional. Instead, this model is intended to be used as a screening tool to assist healthcare professionals identify patients who may need more support and monitoring. By identifying patients at increased risk for hypoglycaemia, additional support and monitoring to prevent or reduce the severity of hypoglycaemic events can be targeted to patients who are most in need of such support. By opting for a relatively low cut-off point, a high sensitivity can be achieved, ensuring that few patients at increased risk of hypoglycaemic events are missed. It should be noted that such a tool is not appropriate for detecting patients at risk of minor hypoglycaemic events that are not reported to GPs.

**Conclusion**

We developed a model for identifying patients at increased risk of hypoglycaemia in primary care using demographic and medication data based on LASSO, which can be used as a screening tool. This model had an acceptable performance, outperformed individual predictor variables and performed similarly to a model that used additional clinical data.

**Acknowledgement**

An unconditional grant was provided by the Royal Dutch Pharmacists Association (KNMP), they had no role in the execution of this study or in the drafting of the article.

**Authors' Contribution Statement**

SC, KT and PD: research idea and study design. SC, PD: Data acquisition. SC, SB, KT and PD: analysis and interpretation. SB, PD and KT: supervision or mentorship. All authors contributed substantially to the intellectual content during manuscript drafting or revision. All authors approved the manuscript and this submission.
CHAPTER 3

Conflicts of interest
The authors have no conflicts of interest.

Availability of data
Requests to receive data from the GIANTT database can be directed at the GIANTT database steering commission https://www.giantt.nl/.
Identifying patients at increased risk of hypoglycaemia in primary care

References


CHAPTER 3


Identifying patients at increased risk of hypoglycaemia in primary care


PART II

Patients’ and healthcare professionals’ perspectives on deprescribing and hypoglycaemia