Chapter 1
General introduction
Epidemiology is defined as “the study of the occurrence and distribution of health-related states or events in specified populations” [1]. In epidemiological research, such health-related states or events are often studied in relation to other factors, or determinants [2]. A common type of epidemiological study is a cohort study. In a cohort study, a group of individuals is followed over time, the determinant or determinants of interest are measured, and the individual’s health state is observed during follow-up [3].

This thesis focuses on prediction models and etiological models in epidemiological cohort studies in two fields: the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia. Type 2 diabetes is a chronic metabolic disorder, characterized by high blood glucose levels. Dementia is a syndrome characterized by a progressive decline in cognitive functions, and pneumonia is a highly prevalent disease and an important cause of death for people with dementia. These two fields might not seem related at first, but for epidemiological research there are several similarities: first, type 2 diabetes and dementia are chronic diseases, characterized by a progressive decline in, respectively, glycaemic control or cognitive functions. Second, type 2 diabetes and dementia are highly prevalent in our current society. In 2015, it was estimated that the global prevalence of diabetes amongst adults was 8.8% (415 million people) [4]; type 2 diabetes is the most common type of diabetes and it is estimated that in high-income countries, about 90% of the people with diabetes have type 2 diabetes [4]. In the same year, it was estimated that almost 47 million people were living with dementia worldwide, which is 5% of all people aged 60 years and over [5]. For these people, pneumonia forms a large burden: for up to 66% of people dying with dementia, pneumonia may be the ultimate cause of death [6-8]. A final similarity between the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia, is that in both disease fields, prediction models can be very relevant for clinical practice.

In the field of pneumonia in nursing home residents with dementia, in this thesis, prediction models are developed, validated and updated for predicting the risk of short-term mortality. These models can be used to inform physicians, patients and their families on prognosis in order to support them in making decisions regarding further treatment [9, 10]. In the field of type 2 diabetes, in this thesis, prediction models are developed, validated and updated for identifying individuals who are at high risk of developing the disease. The use of a prediction model can this way help general practitioners as well as policy makers to target preventive interventions specifically to those high-risk individuals, who are most likely to benefit from the intervention [11, 12]. In addition to prediction modelling, this thesis focusses on the identification of potential etiological factors in the field of type 2 diabetes. The identification of such etiological factors might contribute to understanding the mechanisms that cause type 2 diabetes, which can in turn provide knowledge for possible new interventions to prevent or delay the development of type 2 diabetes or diabetes-related complications [13].
This thesis does not aim to focus on specific mechanisms, but on the methods that are used: prediction modelling and etiological modelling. Also, this thesis does not aim to develop new methodology. Rather, the aim is to apply existing methods for etiological modelling and for prediction modelling to the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia. Prediction modelling is a relatively new field for epidemiologists. Etiological modelling has widely been applied in epidemiology, but this field still faces challenges due to developments such as the increasing availability of data that were routinely collected in clinical practice. The focus of this thesis is on the methodological issues that were encountered in the application of these methods. This way, next to answering research questions that are relevant for the respective disease fields, this thesis may contribute to improvement of these existing methods, by identifying practical challenges and possible research gaps for future research. In addition, this thesis may help researchers to think more critically about how to apply these methods.

This introduction section first provides more information on etiological models and prediction models in epidemiology, and on the difference between those two types of models. Next, more background information is provided regarding the two disease fields that this thesis focusses on, and regarding the cohort studies that are used in this thesis. This way, the research aims of each of the chapters are introduced, as well as the methodological issues that are discussed in the general discussion. Finally, in this introduction section, the main objective and the outline of this thesis are explained.

ETIOLOGICAL MODELS AND PREDICTION MODELS

Two types of epidemiological models are used in this thesis: etiological models, aiming to estimate a causal relationship between a determinant and an outcome; and prediction models, aiming to predict an outcome as accurately as possible. The following part gives some background on these two types of models and on important concepts that are related to these models. Further, important differences between the two model types are highlighted.

Etiological models
In this thesis, etiological models, also referred to as association models, are used to assess etiological relations in the field of type 2 diabetes. This type of model is used in etiological research, which aims to understand the mechanisms that cause a disease or a complication, and thus aims to identify causal relations [13]. Historically, most epidemiological research involved etiological research questions [3]. The origins of etiological research can be found in early research on the causes of communicable diseases. A classic example is the discovery of John Snow in 1855 that the Broad Street pump was the source of a cholera epidemic in London, although in that time, the role of germs as a cause of infectious
diseases was not yet known [3]. Since then, etiological models have widely been used, not only for communicable diseases, but also for other acute diseases as well as chronic diseases, to demonstrate or exclude a causal relation between an exposure and a health outcome [3, 10].

The causal relations examined in etiological models can be disturbed by confounders, which are factors that distort the relation that is being studied [2]. For example, in chapter 7 of this thesis, the association between hypoglycaemia (i.e., episodes of low blood glucose) and mortality is investigated for people with type 2 diabetes. In this relation, diabetes duration might be a confounder: people with a longer diabetes duration might experience hypoglycaemia more often than people with a shorter diabetes duration. In addition, people with a longer diabetes duration also have a higher mortality risk. Thus, an observed association between hypoglycaemia and mortality risk might be explained in part by diabetes duration. In order to 'causally isolate' the relation between the determinant of interest and the outcome, in etiological models, it is important to correct for potential confounding factors. Such corrections are usually performed using multivariable models, which are statistical models including multiple variables. It is important to emphasize, however, that the interest of an etiological model is in the relation of a single determinant with the outcome, while the other variables in the multivariable model are included in order to correct for confounding. As explained in the next part, this should not be confused with multivariable models including multiple determinants of interest, like in prediction models.

**Prediction models**

In this thesis, prediction models are developed, validated and updated in the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia. Whereas etiological research can aid to identify causes of a disease, which might be useful to prevent diseases or to find new treatments, prediction models might help to answer questions from clinical practice regarding adequate diagnosis and regarding prognosis [3]. Epidemiological prediction models can be diagnostic models or prognostic models. Diagnostic prediction models are used to diagnose a specific disease or condition: these models estimate the probability that the outcome is currently present [14]. This thesis focuses on prognostic prediction models. Prognostic prediction models estimate the probability that a certain disease state, or outcome, will occur in the future, based on patient characteristics [10, 14]. In epidemiology, a prognostic prediction model can be called a risk prediction model, risk prediction tool, prognostic model, prognostic index, prediction rule or risk score [14]. Similarly, predictors can be called prognostic factors, risk indicators, determinants, covariates or independent variables [14].

In recent decades, prognostic prediction models have increasingly been used to inform physicians, patients and patients' families on disease prognosis, in order to aid decision-making on disease treatment [9]. An important reason for the increased use of
prediction models is the development of evidence-based medicine: decisions about the care of individual patients are explicitly based on current best evidence [9, 15]. This has led to clinical practice guidelines being increasingly based on evidence from prediction models [14]. Moreover, those clinical practice guidelines increasingly recommend the use of prediction models in clinical practice [14]. In addition to aiding decision-making about disease treatment, prognostic prediction models can be used to inform healthy individuals, physicians and policy makers on the risk of developing a disease in the future, in order to plan lifestyle or therapeutic interventions [10].

To develop a prediction model, potential predictors are combined into a multivariable model, aiming to develop a model that predicts the risk of a future outcome as accurately as possible [10]. Thus, in contrast to etiological models, the interest of a prediction model is not in the relation of one determinant with the outcome, but in finding a combination of determinants that can accurately predict the outcome. Moreover, in contrast to etiological models, causality is not an issue in prediction models: any patient characteristic that is associated with the outcome can be a predictor, whether this association is causal or not [10]. For example, in chapter 3 of this thesis, a model is described that predicts the risk of developing cardiovascular disease, type 2 diabetes or chronic kidney disease. One of the predictors in this model is ‘use of antihypertensive medication’. The use of antihypertensive medication does not cause the development of these diseases. It does indicate that a person has hypertension, and it was found to be an important predictor of these diseases. As predictors are not necessarily causal factors, they should not be used to select or design preventive interventions [13]. This is clear from the above example, where stopping antihypertensive medication would obviously not be a good preventive intervention to reduce the risk of developing cardiovascular disease, type 2 diabetes or chronic kidney disease. Even when a predictor might have a causal relation with the outcome, it is not necessarily a possible factor to intervene on. For example, an important predictor for the development of type 2 diabetes is age: people with a higher age have a higher risk for developing type 2 diabetes. However, reducing age is obviously not a possible intervention to prevent type 2 diabetes. Rather, people who are identified as having a high risk to develop cardiovascular disease, type 2 diabetes or chronic kidney disease might get lifestyle advice and possibly drug treatment [16].

As establishing causal relations is not the aim of prediction models, correcting observed associations for possible confounders is, at least in theory, not an issue in prediction modelling. It has been noted, however, that in practice, considering important confounders might enhance the generalizability of a prediction model [13]. This is related to one of the methodological issues discussed in this thesis: Does confounding only play a role in etiological models?
Prediction modelling: methods for development, validation and updating

Prediction modelling is a relatively new field within epidemiology, and the methods to develop, validate and update prediction models are not yet common knowledge for epidemiologists, so there is little shared understanding. Several reviews have shown that the reporting of prediction models is poor [17-19]. For example, a review of 86 prediction models published in leading general medical journals between 2006 and 2009 showed that less than a quarter of the papers included an external validation study [17]. Another review showed that even in validated tools, more than half did not report calibration performance of the model [19]. In 2009, a series of papers has been published that provides an overview of the principles and methods of prognostic prediction modelling [10, 20-22]. In addition, the TRIPOD statement has been published in 2015, describing guidelines for the reporting of multivariable prediction models [14]. For example, the TRIPOD statement recommends that prediction modelling studies should report discrimination as well as calibration [14]. In the next part, some principles and methods for developing, validating and updating prediction models are explained that are currently recommended and that are applied in this thesis.

When developing a prediction model, candidate predictors can be various factors observed or measured from individuals, such as demographic factors, lifestyle factors, clinical factors, physical measurements, or blood markers [9, 10]. To develop a prediction model that generalizes to practice, predictors should be clearly defined, standardized and reproducible [10]. Furthermore, to facilitate the use in practice, the application of the prediction model should not be too time-consuming and prediction models may focus on predictors that are easy and cheap to measure. Moreover, the use of the prediction model should not lead to burdensome diagnostic tests for the patient, for example, because an X-ray or a blood marker that would otherwise not be determined, is needed to calculate a risk score. Therefore, this thesis focuses on prediction models that include only predictors that can be measured non-invasively, which makes the models feasible for clinical practice or for large-scale prevention programs.

To select the predictors that are relevant to include in the prediction model, stepwise selection is often used, selecting predictors based on their statistical significance [20]. This can be done using a backward selection procedure: first, all candidate predictors are combined into one model –the full model–, and second, the least significant candidate predictor is excluded from the model. The second step is repeated until a pre-specified significance level is reached, often \( p < 0.157 \), according to Akaike's criterion [23]. Alternatively, a forward selection procedure can be used: the model is built up by adding significant candidate predictors step by step, again, until a pre-specified significance level is reached, again, often \( p < 0.157 \). A disadvantage of forward selection, is that it may fail to select correlated predictors, as these may not reach significance independent of the other, thus leading to possibly important predictors not being included in the prediction model. Therefore, in this thesis, backward selection is used to select predictors, which is
preferred over forward selection, as it enables the judgement of all candidate predictors simultaneously [9].

After a prediction model has been developed, several measures can be used to evaluate whether the model can accurately predict the outcome. As recommended by the TRIPOD statement [14], in this thesis, measures for calibration and discrimination are reported. Calibration is the ability of a prediction model to predict absolute risks to develop the outcome. Calibration performance is assessed statistically using Hosmer and Lemeshow goodness-of-fit (H&L) statistics, and visually using calibration graphs. Discrimination is the ability of a prediction model to distinguish between individuals who are at high risk to develop the outcome and those at low risk. Discriminative performance is assessed using area under the receiver operating characteristic curves (AUCs). An AUC of 0.5 would indicate that the model does not perform better than flipping a coin, while an AUC of 1.0 would indicate perfect discrimination. In addition, this thesis reports reclassification tables and variation explained by the model (R²), a measure for overall performance. Reclassification tables can be used to compare classifications of patients as low or high risk between two models. This way it can be evaluated whether the second prediction model improves the way patients are classified as low risk or high risk compared to the first model [24]. For a prediction model with a continuous outcome, this thesis reports the variation explained by the model (R²), which can be considered as an overall measure of the predictive ability of a model [9, 25].

This way, the performance of a prediction model can first be assessed in the development dataset, i.e. the dataset that was used to develop the prediction model. This is called the apparent performance. The apparent performance of a prediction model is often optimistic, due to several factors, including the use of backward or forward selection to select predictors, a small development dataset, and considering a relatively large number of candidate predictors [14]. All these factors, and especially the combination of these factors, can lead to overfitting: the prediction model might be too closely adapted to the dataset that was used to develop the model, limiting the generalizability of the model to other populations, and resulting in optimistic estimates of the apparent performance of the model. However, the extent of optimism can be quantified by performing an internal validation. Internal validation is therefore considered an essential step after developing a prediction model [14]. In this thesis, internal validation procedures are performed using bootstrapping, which is one of the preferred methods for internal validation [14]. Subsequently, the estimated optimism is subtracted from the apparent performance measures, and optimism-corrected performance estimates are reported [9, 14].

Ultimately, a prediction model should provide accurate risk estimations for individuals that were not included in the modelling process, and the model should accurately discriminate between future low risk and high risk individuals. Therefore, after developing and internally validating a prediction model, the next step is to assess its generalizability by performing an external validation [14]. This way, the performance of the model can
be evaluated in a population that was not used in the development of the model. Even when a prediction model is internally validated to correct for optimism, the performance of the model might be suboptimal in the new population, due to differences between the populations. In this thesis, existing prediction models are externally validated in populations that differ from the populations that were used to develop the models in several aspects, including geographical differences, different time periods, and even a different treatment strategy. When the performance of a model is suboptimal in the new population, the model can be updated to improve its performance in the new population. In this thesis, different methods to update prediction models are applied, as described by Steyerberg, including methods for recalibration, methods for model revision and model extension [9]. This is related to another methodological issue discussed in this thesis: Can the same prediction model be applied in different populations, even when a population is treated differently? Is updating needed when a prediction model is applied in a different population?

**DISEASE FIELDS AND RESEARCH AIMS**

These methods for prediction modelling and for etiological modelling are applied to two disease fields: the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia. The following part gives some background information on these diseases, and introduces the research aims that are addressed in this thesis.

**Type 2 diabetes**

Type 2 diabetes mellitus is a chronic metabolic disorder, characterized by high blood glucose levels, or hyperglycaemia, resulting from insulin resistance and/or a progressive insulin secretion defect [26, 27]. These high blood glucose levels increase the risk of diabetes-related microvascular complications: retinopathy, nephropathy, and neuropathy, as well as macrovascular complications: ischaemic heart disease, stroke and peripheral vascular disease [28]. Type 2 diabetes is diagnosed based on fasting glucose levels and/or glucose levels after an oral glucose tolerance test, and since 2010, also based on glycated haemoglobin (HbA1c) levels, which reflect average blood glucose levels over the last 2-3 months [29].

Type 2 diabetes can be prevented or delayed by lifestyle interventions and/or drug interventions [30-32]. These interventions are most successful and most cost-efficient when targeted to individuals at high risk of developing the disease, instead of randomly targeting the general population [33]. As mentioned previously, a prediction model can help to target interventions specifically to those high-risk individuals. Such a prediction model might be restricted to predictors that can be measured non-invasively, to facilitate its use in clinical practice and for research purposes in large databases where blood
assays are not available. Several non-invasive prediction models have been developed for predicting the risk of type 2 diabetes [34], but these models did not include HbA1c levels as a criterion to diagnose type 2 diabetes, while HbA1c has been added as a diagnostic criterion for type 2 diabetes in 2010 [29]. Chapter 2 of this thesis aims to develop a prediction model that predicts HbA1c levels after six years in a general –non-diabetic– population, including non-invasive predictors that have previously shown to be able to predict type 2 diabetes.

Type 2 diabetes shares many modifiable risk factors with cardiovascular disease and chronic kidney disease, such as smoking, hypertension, and being overweight, and therefore, common opportunities for prevention have been suggested [35-37]. Such a joint prevention program targeting shared risk factors will save time and financial resources, compared to developing separate prevention programs for the individual diseases. To improve the cost-effectiveness of such prevention programs, a clear target population is needed [33, 38, 39], and thus, a prediction model can help to identify individuals at risk for these three so-called chronic cardiometabolic diseases. Such a prediction model was previously developed by our research group, including predictors that can be measured non-invasively [37]. Chapter 3 of this thesis aims to externally validate this prediction model in order to evaluate the generalizability of the model.

Next to prediction models in the field of type 2 diabetes, this thesis focusses on etiological models in this field. Several etiological factors have been identified that contribute to the development type 2 diabetes, including age, obesity and lifestyle [40]. Recently, disturbance of the circadian rhythm was identified as another etiological factor for developing type 2 diabetes [41]. For example, shift work and time zone travelling disrupt the circadian rhythm [42], and such a disruption of the circadian rhythm is associated with a higher risk of developing type 2 diabetes [43]. Less is known about the effect of the more chronic disruption of the circadian rhythm: social jetlag. Social jetlag is the discrepancy between our internal circadian clock and social clock: rather than following their internally regulated sleep-wake times, people often use alarm clocks to align their sleep and wake times with social obligations, such as work and school schedules [44, 45]. Chapter 6 of this thesis aims to evaluate the association of social jetlag with cardiometabolic risk factors and type 2 diabetes.

After diagnosis of type 2 diabetes, treatment can include diet advice, oral glucose-lowering medication, and/or insulin. A possible side effect of oral glucose-lowering medication or insulin can be hypoglycaemia, or abnormally low blood glucose levels. An episode of hypoglycaemia can give relatively mild complaints, such as tremor, hunger, sweating, difficulty thinking, but it can also lead to seizure, coma, and even death [46]. Up until now, most research into the association between hypoglycaemia and mortality has focussed on objectively measured hypoglycaemia, based on blood glucose measurements or on hospital records or insurance records. However, blood glucose measurements are not always performed in usual care and not all episodes of hypoglycaemia require medical
assistance. Therefore, self-reported hypoglycaemia might better reflect hypoglycaemia as experienced in everyday life by people with type 2 diabetes. **Chapter 7 of this thesis aims to evaluate whether self-reported hypoglycaemia is associated with short-term mortality in insulin-treated people with type 2 diabetes.**

Good glycaemic control is very important for people with type 2 diabetes in order to prevent or delay diabetes-related complications [40]. It has been shown that a low socio-economic status is associated with poor glycaemic control [47]. This association might be explained by several factors, such as access to care, social support, diabetes-related knowledge, and the ability to adhere to medication, exercise and dietary regimens [47]. In this thesis, it is hypothesized that also the socio-economic status of their partner might influence the individual’s level of glycaemic control. **Chapter 8 of this thesis aims to evaluate the association between individual and partner’s socio-economic status and HbA1c levels in people with type 2 diabetes.**

**Pneumonia in nursing home residents with dementia**

Dementia is a syndrome characterized by a progressive decline in cognitive functions [48]. In the Netherlands, over 40% of the nursing home residents suffer from dementia [49-51]. Dementia is associated with a reduced life expectancy [52], and palliative care is often seen as appropriate for people with dementia, which means that the focus is on improving comfort, by treating pain and other symptoms, while preventing overtreatment with possibly non-beneficial and burdensome treatment [53, 54]. A highly prevalent [55] and burdensome disease in nursing home residents with dementia is pneumonia: it is associated with discomfort and is an important cause of death [6-8, 56, 57]. Although pneumonia is normally treated with antibiotics, it may sometimes be questioned whether or not treatment with antibiotics is desirable for patients with dementia. For example, it is not clear whether antibiotics actually improve comfort [58, 59]. Moreover, in people with advanced dementia, it has been suggested that antibiotics may prolong the dying process, rather than extending life [60, 61]. In addition, patients’ wishes regarding treatments are often unclear [62]. Therefore, for physicians and families involved in the care of people with dementia it may be difficult to decide how to treat the patient, e.g., whether to try to cure the pneumonia with antibiotic treatment, or to prepare for a dying process and focus on comfort [63]. A prediction model can inform on prognosis and may this way support decision-making. Such a prediction model was previously developed by our research group, predicting the risk of short-term mortality for nursing home residents with dementia and pneumonia treated with antibiotics [64]. **Chapter 4 of this thesis aims to predict short-term mortality for residents not treated with antibiotics.** It was previously shown that predictors for short-term mortality were largely similar between those two groups [65]. Therefore, rather than developing a new prediction model for the untreated residents, this chapter aims to evaluate whether the model that was previously developed for treated residents could be applied to residents who were not treated with antibiotics.
An additional goal is to evaluate whether updating of the model is needed to improve model performance in the untreated residents, and to externally validate this updated model.

The above-mentioned model for predicting short-term mortality for residents with dementia and pneumonia treated with antibiotics was developed in the 1990s. Since then, the care for older people has changed as well as resident characteristics. For example, the prevalence of malnutrition in nursing home residents has decreased over the years [49], cardiovascular risk management has improved in the general population as well as in older people [66, 67], and new policies promote people to be cared for at home as long as possible [68]. Therefore, chapter 5 of this thesis aims to evaluate the performance of the previously developed model in a more recent cohort study (2012-2015). In addition, this chapter aims to evaluate whether the performance can be improved by revising predictors and/or adding additional predictors to the model, specifically, predictors related to nutrition, hydration, overall health condition, and cardiovascular risk.

COHORT STUDIES

To address the research aims that are introduced above, 16 different cohort studies are used in this thesis, which are described in more detail in the relevant chapters. In the following part, three types of cohort studies are introduced: 1) population-based cohorts, 2) nursing home cohorts, and 3) cohorts using data that were routinely collected in clinical practice.

In several chapters of this thesis, individual patient data from different cohort studies are combined, which is also referred to as pooled analysis. This can be expected to increase the generalizability of the results compared to the analysis of a single study: the results are derived using a broader population and may thus apply to a broader population [69, 70]. However, cohort studies may use different definitions and availability of variables may differ between the cohorts. This is related to another methodological issue discussed in this thesis: How can etiological and prediction models be developed using combined data from different cohort studies?

Population-based cohorts

In chapters 2, 3 and 6 of this thesis, 8 different population-based cohorts are used to develop two prediction models and one etiological model in the field of type 2 diabetes: The Hoorn Study (The Netherlands) [71], the Inter99 Study (Denmark) [72], the Cooperative Health Research in the Region of Augsburg (Germany) [73], the Metabolic Syndrome in Men Study (Finland) [74], The Australian Diabetes, Obesity and Lifestyle Study (Australia) [75], The Rotterdam Study (The Netherlands) [76], The Prevention of Renal and Vascular End-stage Disease study (The Netherlands) [77], and the New Hoorn Study (The Netherlands) [71]. These populations were arrived at by inviting a random sample of all residents of a
specific city or region for research. The populations in these cohort studies were observed and followed up over time in order to study relationships between determinants and outcomes [78]. Participants underwent physical examinations, blood samples were taken, and/or participants filled in several questionnaires.

**Nursing home cohorts**

In chapter 4 and 5 of this thesis, 6 nursing home cohorts are used: The Dutch Pneumonia Study (The Netherlands) [63], the Missouri Lower Respiratory Infection Study (USA) [79], The ‘Dutch 2006–2007’ study (The Netherlands) [80], The ‘Bedford US’ study (USA) [61], and the Dutch End of Life in Dementia study (The Netherlands) [81], and the PneuMonitor study (The Netherlands) [82]. These cohort studies were set up to study pneumonia – or lower respiratory infections– and/or end-of-life care. In contrast with the above-mentioned population-based cohorts, specific populations were included in these cohorts, namely nursing home residents, mostly residents with dementia. In these cohort studies, the attending physicians or research nurses completed questionnaires regarding the residents.

**Cohorts using data that were routinely collected in clinical practice**

Finally, in chapter 7 and 8 of this thesis, two cohorts of people with type 2 diabetes are used that consist of data that were routinely collected in clinical practice: the Hoorn Diabetes Care System Cohort (The Netherlands) [83] and the Diabetes Pearl cohort (The Netherlands) [84]. Whereas in the above-mentioned cohorts, physical examinations were undertaken and questionnaires were filled in for the purpose of the study, in these two cohorts, data were –at least partly– collected as part of routine clinical care and were retrieved from hospital information systems. The use of data that were routinely collected in clinical practice has several advantages for epidemiological research: routinely collected patient data are increasingly getting available and this way, available data of real-life care are used rather than needing to set up a cohort, which is much more expensive [85]. In addition, included patients are likely to reflect clinical practice regarding the availability of variables and the quality of the data [3], and regarding the patients characteristics. However, there are also several challenges when using routinely collected patient data for research purposes, related to checking the quality of the data and cleaning the data. This is related to another methodological issue discussed in this thesis: How can data that are routinely collected in clinical care be used for research purposes?

**MAIN OBJECTIVE**

The main objective of this thesis is to develop, validate and update prediction models and to identify potential etiological factors in epidemiological cohort studies in the
field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia. This thesis aims to apply existing methods for prediction modelling and for etiological modelling to these fields. The focus of this thesis is on the methodological issues encountered in the application of these methods, including: How can etiological and prediction models be developed using combined data from different cohort studies? Does confounding only play a role in etiological models? How can data that are routinely collected in clinical care be used for research purposes? Can the same prediction model be applied in different populations, even when a population is treated differently? Is updating needed when a prediction model is applied in a different population?

OUTLINE THESIS

In chapters 2-5, prediction models in the field of type 2 diabetes and the field of pneumonia in nursing home residents with dementia are developed, validated and updated. In chapter 2, a prediction model is developed and validated that aims to predict HbA1c levels after six years in the general population. In chapter 3, an existing risk prediction model for predicting the combined 7-year risk of developing type 2 diabetes, cardiovascular disease or chronic kidney disease in the general population is externally validated. In chapter 4, a prediction model that was previously developed for predicting mortality for nursing home residents with dementia and pneumonia treated with antibiotics is validated and updated for residents who were not treated with antibiotics. In chapter 5, the same prediction model for treated residents is validated and updated by revising predictors and extending the model with new predictors in a more recent population.

In chapters 6-8, etiological relations in the field of type 2 diabetes are assessed. Chapter 6 evaluates whether social jetlag –which is the discrepancy between our internal circadian clock and social clock– is associated with a higher prevalence of cardiometabolic risk factors or a higher prevalence of type 2 diabetes in the general population. Chapter 7 evaluates whether self-reported hypoglycaemia is associated with the risk of mortality in people with type 2 diabetes. Chapter 8 evaluates whether the level of occupation of people with type 2 diabetes and their partners is associated with HbA1c levels in those people.

Finally, chapter 9 is a general discussion of the results described in this thesis, reflecting on the main findings, discussing methodological issues, and providing implications and recommendations for practice and for further research.
REFERENCES


50. Sociaal en Cultureel Planbureau (SCP), Ouderen in verpleeghuizen en verzorgingshuizen - Landelijk overzicht van hun leefomstandigheden in 2015/16 [In Dutch]. 2017: SCP.


63. van der Steen JT, Ooms ME, Adèr HJ, Ribbe MW, and van der Wal G, *Withholding antibiotic treatment in pneumonia patients with dementia: A quantitative observational study*. Archives of Internal Medicine, 2002; 162(15):1753-60.

64. van der Steen JT, et al., *Predictors of mortality for lower respiratory infections in nursing home residents with dementia were validated transnationally*. Journal of Clinical Epidemiology, 2006; 59(9):970-79.


Chapter 1


26